Mémoire de Maîtrise universitaire interfacultaire en humanités numériques

LA RECONNAISSANCE AUTOMATIQUE DE L'ÉCRITURE MANUSCRITE ET LES CAHIERS DE JEAN-HENRI POLIER DE VERNAND (1715-1791)

Présenté dans la discipline Histoire

Par

Titaÿna Kauffmann

sous la direction du Professeur Béla Kapossy
et la codirection de Lucas Arnaud André Rappo

Session de Printemps 2023

Remerciements

Je tiens à remercier tout mon entourage qui m'a soutenu dans la réalisation de ce mémoire. Je remercie également mes co-directeurs de mémoire, Béla Kapossy et Lucas Arnaud André Rappo, qui ont su m'aider à orienter mon travail. Je remercie également Rémi Petitpierre pour m'avoir montré comment utiliser HTR-Flor++ et aidé à rédiger les scripts concernant la reconnaissance automatique de texte

TABLE DES MATIÈRES

PLAN

LISTE DES ABRÉVIATIONS

ACV	Archives Cantonales Vaudoises
CER	Character Error Rate
CRNN	Convolutional Recurrent Neural Networks
HMM	Hidden Markov Model
HTR	Handwritten Text Recognition
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NLP	Natural Language Processing
OCR	Optical Caracter Recognition
WER	Word Error Rate

1. INTRODUCTION

1.1. PROBLÉMATIQUE

Dès 2021, un projet d'océrisation (OCR) des cahiers de Jean-Henri Polier de Vernand se fait en collaboration entre les Archives cantonales vaudoises (ACV) et le Collège des Humanités digitales (CDH) de l'EPFL[1]. Après avoir fourni les documents numérisés au format PDF, il s'agira dans ce travail de mettre en place une reconnaissance automatique de texte permettant d'extraire de la source physique un texte au format informatique.

Jean-Henri Polier de Vernand est né en 1715 à Lausanne et y occupa dès 1754 jusqu'à sa mort en 1791 le rôle de lieutenant baillival de Lausanne et donc remplaçant du bailli bernois[2]. Membre de deux Conseils de Lausanne, de la Cour baillivale, de la Cour des fiefs, de la Cour criminelle du Château, de la Cour du Chapitre et membre du Tribunal de la Rue de Bourg grâce à son statut de propriétaire, il était l'un des personnages les plus importants de la vie lausannoise de son époque[3]. Qui plus est, Polier s'est attelé à tenir de nombreux cahiers dès 1754 jusqu'à sa mort où il a méthodiquement retranscrit ses journées sur plus de 26 300 pages, faisant de cette source si particulière l'un des «plus importants documents d'histoire lausannoise»[4].

Ce genre d'écrits personnels à l'image des cahiers de Jean-Henri Polier de Vernand connaissent un nouvel intérêt en histoire en Suisse romande avec tout ce que l'on peut réunir «sous le vocable d'"egodocuments", "d'écrits du for privé", ou de littérature de témoignage, voire d'"auto-témoignage" — terme le plus proche de celui généralement retenu par l'historiographie germanophone de "Selbstzeugnis"».[5] Ce qui rassemble ces différents écrits sont notamment le caractère autobiographique tout comme la capacité d'expression de la «variabilité des expressions de soi dans les divers contextes historiques et sociaux»[6]. Ainsi il

---

[1] ACV, « Archives cantonales vaudoises : rapport d'activités 2021 by Etat de VAUD - Issuu », 12.08.2022, https://issuu.com/etatdevaud/docs/rapport-annuel-2021, consulté le 06.03.2023.

[2] Abetel Emmanuel, « Polier de Vernand, Jean-Henri », *hls-dhs-dss.ch*, https://hls-dhs-dss.ch/articles/017839/2009-04-20/, consulté le 03.03.2023.

[3] Morren Pierre, *La vie lausannoise au XVIIIe siècle: d'après Jean-Henri Polier de Vernand, lieutenant baillival*, Genève : Labor et Fides, 1970, préface I

[4] Abetel Emmanuel, « Polier de Vernand, Jean-Henri »…, *art. cit.*

[5] Tosato-Rigo Danièle, « Pratiques de l'écrit et histoire par la marge : autour des "egodocuments" en Suisse romande au XVIIIe siècle », Verlag Karl Schwegler AG, 2010, p. 261

[6] *Ibid.*

s'agira notamment ici de relier les écrits de Polier à leur contexte de la ville de Lausanne au XVIIIe siècle.

La position particulière qu'occupe le lieutenant vis-à-vis des autorités bernoises Berne durant le dernier siècle de l'ancien régime, mais aussi de la rigueur avec laquelle il a rempli les pages de son Livre de Raison fait dire à Morren que le titre de «Mémorial» est plus approprié, celui-ci dépassant le cadre familial généralement associé au premier[7].

En ce sens, Jean Polier peut être considéré comme un cas-limite — représentatif soit typiquement ou atypiquement —, au sens où l'entendait l'historien italien Carlo Ginzburg :

> Soit négativement — car il aide à préciser ce qu'il faut entendre, dans une situation donnée par "statistiquement le plus fréquent". Soit positivement — car il permet de circonscrire les possibilités latentes de quelque chose qui ne nous est connu qu'à travers une documentation fragmentaire et déformée.[8]

Dans *Le fromage et les vers*, Ginzburg s'attache à écrire la biographie d'un meunier au XVIe siècle en élargissant «vers le bas le concept historique "d'individu"»[9]. S'il s'agit là d'un cas typique, Ginzburg définit la biographie comme méthode historique dont la démarche est de «démêler les multiples fils qui lient un individu à un milieu et à une société historique»[10] tout en ayant des enjeux propres. Ici, l'on se trouve dans une situation atypique qui nous permettrait d'explorer le champ de la position particulière occupée par le lieutenant.

Au moyen de la reconnaissance automatique d'écriture manuscrite, ce projet entend dans un premier temps rendre disponible dans un format numérique l'intégralité des 26 300 pages rédigées par Jean-Henri Polier de Vernand. Ainsi celui-ci serait mis à disposition des chercheurs — sur un support encore non déterminé à ce stade — dans son intégralité et en restant au plus proche de la source et de la manière dont le Lausannois prit le soin de rédiger au jour le jour ses cahiers.

En effet, si de nombreux écrits personnels romands allant du XVIIIe au début du XXe siècle sont d'ores et déjà publiés, la publication relève généralement d'initiatives privées où les «descendants de familles de notables héritiers d'un riche patrimoine documentaire ont largement puisé dans les écrits personnels de leurs ancêtres pour leur rendre hommage et

---

7 Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 10
8 Ginzburg Carlo, *Le fromage et les vers: l'univers d'un meunier du XVIe siècle*, Paris : Flammarion, 1980, p. 16
9 *Ibid.*
10 *Ibid*.

composer quelques tableaux de la vie locale»[11]. Dès lors les textes publiés font l'objet d'un travail éditorial, que ce soit par les descendants qui appliquent une censure basée sur des critères moraux ou encore par les historiens qui visent à effectuer des sélections thématiques qui relèvent parfois de l'anachronisme.

L'historienne spécialiste de la culture de l'écrit Danièle Tosato-Rigo cite notamment l'exemple de Pierre Morren, auteur d'un livre portant sur l'histoire lausannoise au XVIIIe siècle basé sur les écrits de Polier — qui, on le verra dans la description du corpus de source est lié à la famille ayant héritée des cahiers[12]. L'historienne insiste également sur la nécessité de rendre disponible aux chercheurs ce texte souvent cité, mais difficilement accessible[13].

Ainsi dans ce travail, nous reviendrons sur comment les avancées en matière de *machine learning* permettent de l'automatisation de la reconnaissance d'écriture manuscrite afin de rendre disponible l'œuvre de Jean-Henri Polier de Vernand dans son intégralité et son format de rédaction original. Qui plus est, nous nous attarderons également sur les implications qu'une telle technologie peut avoir sur l'histoire et les pratiques archivistiques, tout en détaillant l'état de la recherche sur celle-ci ainsi que son fonctionnement.

Si nous avions au début de ce travail espéré pouvoir utiliser ces résultats pour approfondir l'analyse historique autour de la vie du lieutenant baillival et sa position si particulière au sein du microcosme lausannois durant l'Ancien régime, nous verrons également comment les difficultés techniques qui ont découlé d'un tel projet nous ont amenés à limiter nos ambitions. En effet, les différentes étapes techniques vouées à rendre disponible cette importante source de l'histoire de Lausanne n'ont pas été de tout repos et nous ont forcés à revoir les limites qu'un tel travail en humanités numériques pose quand l'histoire rencontre l'informatique. Ceci dit, il nous faudra d'abord parler de la personne de Jean-Henri Polier de Vernand, sa généalogie, sa formation et son parcours ainsi que sa position sociale, politique et administrative.

1.2. DESCRIPTION DE LA SOURCE

Il semblerait que les archives n'aient jamais quitté Lausanne. À la mort de Jean-Henri Polier de Vernand, ses cahiers sont confiés à son frère, Georges Louis qui les transmit par la suite à son filleul Henri Etienne Polier de Bottens. Par la suite, les cahiers sont transmis à sa

[11] Tosato-Rigo Danièle, « Pratiques de l'écrit et histoire par la marge…, *art., cit.*, p. 261
[12] *Ibid.*
[13] *Ibid.,* p. 265

seconde fille, Juliette Jeanne Pauline de Polier qui épousa Henry de Blonay en 1806. De cette union naît entre autres l'héritier des cahiers, Godefroy de Blonay qui transmit ensuite les documents à sa fille Blanche Alexandrine, née en 1843. Celle-ci épouse en 1867 Gustave Monod et c'est leur fils René Monod qui en héritera avant de déposer les différents cahiers au sein des archives cantonales vaudoises, constituant ainsi le fond du même nom.[14]

Danièle Tosato-Rigo relève notamment que «le caractère justificatif du livre de raison — son scripteur pouvait être amené à le produire en conseil, devant la famille ou des partenaires économiques — et sa fonction de "fil rouge de la lignée" lui ont assuré une plus grande pérennité»[15].

Au sein du fonds privé René Monod, on retrouve 139 cahiers et 21 feuillets d'extraits et de brouillons qui forment les 26 300 pages qui composent notre source, chacun de ces cahiers mesurant 22,5 cm sur 26,5cm et contenant individuellement environ cinquante feuillets reliés non paginés.[16]

Jean-Henri Polier de Vernand a tenu ses différents cahiers dès sa nomination en tant que lieutenant baillival jusqu'à sa mort — soit durant trente-sept ans — en rédigeant chaque jour entre quatre et vingt pages. Cependant, tous les cahiers ne sont pas parvenus jusqu'à nous. Au total, seules dix-sept années comptent l'intégralité des cahiers rédigée par le lieutenant baillival. Les cahiers manquants représentent en général deux à trois mois, à l'exception de l'année 1758 qui n'est malheureusement pas arrivée jusqu'à nous.

Les cahiers vont de 1754 à 1791 au moment du décès de leur auteur. Ils commencent au mois de mai 1754 et se poursuivent jusqu'au mois de septembre 1756. Ils reprennent ensuite en janvier 1757 avec un arrêt durant le mois d'avril jusqu'à la moitié du mois de mai de cette même année. Il y également des manques pour l'année 1757 où dès le mois de juin jusqu'à la fin du mois de septembre sont absents. Le dernier manque intervient au cours de l'année 1759 à la mi-juin jusqu'au commencement du mois de septembre. La dernière anomalie du corpus de source vient de Jean-Henri Polier de Vernand lui-même puisque si les cahiers suivent normalement un ordre chronologique, les premières pages du second cahier qui devraient selon cette logique correspondre à la première partie du mois de février 1755 sont consacrées à la période allant du premier mars 1754 jusqu'au 11 de ce même mois.

---

[14] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 13

[15] Tosato-Rigo Danièle, « Pratiques de l'écrit et histoire par la marge…, *art. cit.*, p.. 261

[16] Favez Valérie, *Etude du « Mémorial universel » tenu par Jean-Henri Polier de Vernand, lieutenant baillival: gestion d'un patrimoine (1754 à 1761)*, Mémoire de Master, Université de Lausanne, 1991, p. 2

Concernant le titre à donner à cette collection de cahiers, Jean-Henri Polier de Vernand utilisait différentes appellations pour ses cahiers, parlant parfois de Livre de raison ou de «Mémorial universel des recettes, dépenses, prêts, emprunts, marchandises prises à crédits et à (?) saison»[17]. À ce titre, Pierre Morren choisit de garder le titre de Mémorial, car celui-ci dépasse largement celui d'un registre de comptabilité d'un foyer, mais est plutôt «composé de remarques et d'observations qu'un personnage ayant une certaine position sociale fait sur les événements marquants de leur vie quotidienne, sur la famille et sur ses comptes»[18]. Suivant cette logique, nous utiliserons également le terme de Mémorial pour désigner ce corpus de source.

Tous les jours Polier de Vernand s'attelait à écrire au sein de ses cahiers le résumé de sa journée, y compris le résumé des séances auxquelles il avait participé ou encore les brouillons des différentes correspondances qu'il entretenait. Qui plus est, Pierre Morren relève que «certains indices fort sérieux nous font croire que l'immense journal qui nous est parvenu est la mise au net d'un brouillon préalable, ce qui doublerait le nombre de pages qu'il rédigea, car jamais il n'utilisa le secours d'un secrétaire»[19].

Ainsi, son emploi du temps hebdomadaire peut se résumer ainsi : le lundi était le jour de la Cour baillivale et des séances du conseil des Soixante et des Deux Cents; en général, le mardi était consacré à ses propres affaires et à des conférences; le mercredi il avait des réunions avec le bailli au sein du Château; le jeudi était le jour des séances de la chambre des orphelins et de celles du consistoire; le vendredi était consacré aux conférences, aux réunions des commissions, à la correspondance officielle et, s'il disposait du temps nécessaire, il s'occupait également de ses propres affaires, tout comme le samedi; le dimanche et les autres jours de la semaine dans des proportions changeantes étaient consacrés à signer ou sceller des mandats de comparution, examiner les dossiers souvent volumineux des affaires qu'il devait juger[20].

Ainsi au fil des 26 300 pages de ses cahiers, Jean-Henri Polier de Vernand relate tout autant sa vie de société — bien qu'il utilise souvent des abréviations lorsqu'ils visitent des personnes connues de la vie sociétale lausannoise — que son activité au Château ou au Conseil de la ville. Il y prend des notes relatives aux différents mandats officiels qu'il sera

[17] P René Morod, 1
[18] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 10
[19] *Ibid*., p. 41
[20] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 41-42

amené à occuper en raison de sa position de lieutenant baillival, à sa correspondance diverse ainsi que toutes ses dépenses et ses créances. Il résume également les sermons religieux suite à ses nombreux passages à l'église.

Il y utilise également son mémorial pour des notes viticoles et agricoles en rapport avec le domaine familial qui lui procure sa principale source de revenus. Un aspect qui pourrait également intéresser d'autres chercheurs – certains peut être moins habitués aux méthodes historiques et qui pourraient bénéficier de la transcription numériques des « cahiers Polier » – est à trouver dans les notes rigoureuses du lieutenant baillival sur ses différents relevés météorologiques quotidiens. En plus des informations sur les récoltes annuelles, ce dernier y inscrit consciencieusement la météo du jour, de la pression barométrique à la température Réaumur, la pluie ou le soleil et la direction du vent.

2. LA RECONNAISSANCE AUTOMATIQUE D'ÉCRITURE MANUSCRITE

2.1. POSSIBILITÉS POUR L'HISTOIRE

Depuis l'avènement de la révolution du numérique, l'un des enjeux épistémologiques majeurs posés aux sciences humaines et sociales et celui de la masse des données désormais disponibles. On se trouve ainsi face à de nouveaux outils et moyens d'observer les pratiques sociales, mais également dans une situation où «les disciplines qui connaissent le mieux le traitement des données numériques se sentent légitimes à étudier le social au même titre que tout autre type de relation, ce qui laisse supposer que le problème principal des sciences sociales est l'absence de données»[21]. Ainsi, les humanités numériques apparaissent comme un champ spécifique d'études se situant à l'intersection entre les pratiques numériques et les sciences humaines et sociales.

Dans ce contexte, il est important que les historiens s'impliquent activement dans la création d'outils de recherches faisant appel aux technologies du numérique adaptées à leurs objectifs de recherche, puisque dans le cas contraire ils risquent de devoir se fier à des outils qui sont souvent conçus en fonction d'ensembles de données non historiques et de questions de recherches non historiques[22].

C'est le cas ici avec la reconnaissance automatique d'écriture manuscrite sur les origines de laquelle nous reviendrons ci-dessous. Pour l'instant, il est intéressant d'observer comment les technologies du numérique ont influé sur les méthodes historiques. Aux fondements de l'histoire, on retrouve les sources, généralement conservées dans des institutions d'archives. Ainsi, les archives ont pendant longtemps été avant tout physique :

> Any of us who have spent time in actual nineteenth-century archives know the literal truth of Jacques Derrida's phrase "archive fever'. As Carolyn Steedman has argued, real archives may well produce something pathological in the researcher that might be named archive fever, because archives reify the period they record. They contain not only the records of a period but its artifacts as well, their dust the debris of toxins and chemicals and disease that went into making the paper and glue and inks, that went into processing the animal skins that wrap the books we open and, in the dusty light, read and inhale.

---

[21] Beaude Boris, « (re)Médiations numériques et perturbations des sciences sociales contemporaines », *Sociologie et sociétés*, vol. 49, nº 2, Les Presses de l'Université de Montréal, 2017, p. 94

[22] Edelstein Dan, « Intellectual History and Digital Humanities », *Modern Intellectual History*, vol. 13, nº 1, Cambridge University Press, 2016, p. 244

> When we emerge from an archive, we are physically and mentally altered. We emerge with notes – photocopies if we're allowed – but never with the archive, which remains behind, isolated from us.[23]

Ed Folsom, historien spécialiste du XIXe siècle et des humanités numériques, met en avant le processus de transformation numérique des archives qui passent de documents physiques à des sources informatisées. Ce passage implique de pouvoir dépasser non seulement la matérialité des archives situées généralement dans un lieu clos où leurs accès sont limités par une demande spécifique adressée à l'archiviste, mais également le développement des bases de données qui facilite l'accès immédiat aux informations et à la juxtaposition des éléments qui dans l'espace réel peuvent se situer très éloigné les uns des autres[24].

Ainsi avec les technologies du numérique non seulement l'accès et la possibilité de faire des liens entre informations est facilité, mais l'on est passé de la figure de l'historien passant des heures dans les archives à prendre des notes manuscrites sur les sources à l'image de plus en plus fréquente en salle de consultation de l'historien en train de prendre des photographies des sources pour pouvoir les emporter ensuite avec lui. Désormais, avec l'arrivée de technologies comme la reconnaissance automatique d'écriture manuscrite, un pas supplémentaire est franchi avec la transformation des sources physiques en fichiers informatiques permettant non seulement un possible accès dématérialisé, mais également un traitement secondaire de la source au moyen de méthodes computationnelles.

On constate donc une évolution vis-à-vis de l'impact des technologies du numérique sur le travail des historiens avec dans un premier temps la possibilité d'identifier les documents pertinents grâce aux nouveaux outils de recherche, puis la possibilité de faire défiler des images scannées des sources dématérialisées des institutions officielles qui les possèdent, voire même désormais de faire des recherches de plein texte par mot-clé dans des ensembles de documents convertis en texte clair[25].

Qui plus est, des initiatives comme celle du *Text Encoding Initiative* (TEI) — qui reflète notamment l'intérêt des sciences humaines et sociales pour l'informatique et les débats des années 1990 autour des textes électroniques et des archives — ont permis la mise en place d'une norme de balisage théorisant les manières de représenter, d'analyser et de diffuser les

---

[23] Folsom Ed, « Database as Genre: The Epic Transformation of Archives », *Pmla-publications of The Modern Language Association of America*, vol. 122, 2007, p. 1577

[24] *Ibid.*

[25] Klein Lauren et Eisenstein Jacob, « Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives », *Scholarly and Research Communication*, vol. 4, no 3, 2013, p. 2

textes électroniques[26]. Ces enjeux redeviennent d'actualité puisque dans les années 1990 on parlait généralement de la publication numérique d'œuvres classiques et désormais avec la reconnaissance automatique d'écriture manuscrite, on peut rendre numériques des sources qui n'avaient jamais été éditées auparavant.

Les documents numériques relevant des normes TEI sont par ailleurs publiés en XML (*Extensible Markup Language*) qui fut publié pour la première à l'occasion d*u World Wide Web Consortium* en 1998, mais qui puise «ses origines dans les systèmes de préparation des documents des années 1980»[27]. Or, le XML est également lié au développement de la reconnaissance automatique d'écriture manuscrite puisque non seulement il présente un format lisible autant pour la machine que pour l'humain, ou encore permet également de lier des images grâce à des informations de mise en page basées sur des pixels qui indiquent des polygones au niveau des régions de texte et des lignes[28].

Étant donné que les collections d'archives sont numérisées à un rythme croissant, les interfaces de recherches dans les bases de données des institutions nécessitant de connaître les termes précis de la recherche ne suffiront plus et les chercheurs auront besoin de nouvelles techniques pour passer au crible et donner un sens à cette quantité croissante de matériel[29]. C'est dans cette logique qu'intervient le *Topic Modeling,* un technique d'exploration de texte qui applique l'inférence probabiliste pour identifier les thèmes latents, ou "sujets", dans un ensemble de documents : «however, due to the difficulty of visualizing a probability distribution over thousands of words, topics are usually summarized by a list of the words with the highest probability relative to other topics»[30].

De manière générale, les modèles de *Machine Learning* se sont considérablement améliorés dans leur capacité à prédire des résultats, y compris dans les corpus de texte non ou semi-structurés. Ainsi les approches informatiques peuvent effectuer par elles-mêmes des déductions extrêmement précises et les méthodes d'apprentissage automatiques peuvent

---

26 Hockey Susan, « The History of Humanities Computing », in *A Companion to Digital Humanities*, John Wiley & Sons, Ltd, 2004

27 Burnard Lou, « La TEI et le XML », in *Qu'est-ce que la Text Encoding Initiative ?*, Marseille : OpenEdition Press, 2015, http://books.openedition.org/oep/1298.

28 Purcell Jake, « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », vol. 7, nº 0, Ubiquity Press, 2021, p. 3

29 Klein Lauren et Eisenstein Jacob,« Reading Thomas Jefferson with TopicViz…, *art cit.* p. 2

30 *Ibid.*

agir comme des extensions de notre capacité cognitive et ainsi aider considérablement à la recherche[31].

En effet, la capacité de mémoire de la machine permet d'augmenter les limites d'analyse de l'homme tout en y incorporant une multitude de caractéristiques linguistiques simultanément afin de pouvoir les associer dans une comparaison constante et fiable[32]. Il est certes important de noter que les méthodes informatiques ne peuvent pas remplacer un chercheur qualifié, mais qu'elles peuvent mettre en évidence de nouvelles données comme par exemple les régularités, les associations et les structures d'un corpus de texte toujours plus important, permettant ainsi aux sciences humaines et sociales d'exploiter ces mêmes méthodes pour affiner et élargir les déductions de la recherche sur le monde social qui sont sous-jacent à la communication, y compris l'écriture[33].

On peut donc également mettre en avant que l'avancée des technologies du numérique pourrait avoir pour effet d'utiliser des méthodes quantitatives pour améliorer les méthodes qualitatives :

> Nevertheless, we show that recent advances in NLP[34] and ML[35] are being used to enhance qualitative analysis in two ways. First, supervised ML prediction tools can "learn" and reliably extend many sociologically interesting textual classifications to massive text samples far beyond human capacity to read, curate, and code. Second, unsupervised ML approaches can "discover" unnoticed, surprising regularities in these massive samples of text that may merit sociological consideration and theorization.[36]

On peut ainsi en histoire faire des liens entre l'analyse de longue durée telle que théorisée par Ferdinand Braudel de l'École des Annales et celui de *Big Data*, puisque des auteurs comme Jo Guldi et David Hermitage théorisent dans *The History Manifesto* qu'en combinant ces deux concepts qui peuvent tous deux recouper l'idée de l'exploration d'un grand nombre de données sur de longues périodes de temps permettent à l'historien d'élaborer de plus

---

31 Evans James A. et Aceves Pedro, « Machine Translation: Mining Text for Social Theory », *Annual Review of Sociology*, vol. 42, nº 1, 2016, p. 23
32 *Ibid.,* p. 25
33 *Ibid.*
34 Natural Language Processing
35 Machine Learning
36 *Ibid.,* p. 22

grands et plus précis récits sur le monde social, permettant ainsi de remplir le rôle social de l'historien[37].

2.2. DÉVELOPPEMENT ET FONCTIONNEMENT DE L'HTR

La reconnaissance automatique de texte manuscrit — *Handwritten Text Recognition* (HTR) — est un défi de recherche majeure depuis plusieurs décennies. Il permet la retranscription de texte cursif écrit à la main en format numérique (ASCII, Unicode)[38]. À l'origine, il s'est particulièrement développé dans le monde de la finance et du commerce avec par exemple l'interprétation des adresses postales, la reconnaissance des chèques bancaires ou encore la reconnaissance de signature[39]. La diversité des documents — pouvant aller entre autres des manuscrits historiques aux prescriptions médicales ou encore des formulaires — souligne la nécessité de construire des modèles de reconnaissance automatique de texte qui puissent être applicables à large échelle[40].

Depuis le milieu du XXe siècle, la reconnaissance automatique de texte est un domaine de recherches des sciences informatiques et a commencé avec le développement de la reconnaissance optique de caractères — *Optical Character Recognition* (OCR) — dans laquelle les images scannées de textes imprimés sont converties en texte codé par une machine, généralement en comparant des caractères individuels à des modèles existants[41]. L'OCR est basée sur le modèle statistique des Markov cachés — Hidden Markov model (HMM) — qui fait partie d'une famille d'outils modélisant des processus séquentiels et largement utilisés dans les étapes de prétraitement et de reconnaissance de texte[42].

Avec les avancements du *Deep Learning* depuis les années 2010 et ceux de l'architecture de réseau neuronal à partir de cellules appelées à long court terme — *Long Short-Term Memory* (LSTM) — la reconnaissance automatique d'écriture manuscrite a été rendue

---

37 Guldi Jo et Armitage David, *The History Manifesto*, Cambridge : Cambridge University Press, 2014, p. 80-81

38 Sousa Neto Arthur Flor de *et alii*, « HTR-Flor++: A Handwritten Text Recognition System Based on a Pipeline of Optical and Language Models », in *Proceedings of the ACM Symposium on Document Engineering 2020*, New York, NY, USA : Association for Computing Machinery, 2020, p. 1

39 Muehlberger Guenter *et alii*, « Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study », *Journal of Documentation*, vol. 75, no 5, Emerald Publishing Limited, 2019, p. 956

40 Sousa Neto Arthur Flor de et alii, « HTR-Flor++: A Handwritten Text Recognition System… », art. cit., p. 1

41 Muehlberger Guenter *et alii*, « Transforming scholarship… », *art. cit.,* p. 995

42 Nockels Joe *et alii*, « Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research », *Archival Science*, vol. 22, no 3, 2022, p. 368

possible pour les sciences humaines et sociales[43]. En effet, le *Deep Learning* est basé sur l'utilisation de ce qu'on appelle des réseaux neuronaux profonds — inspirés des réseaux neuronaux des humains et des animaux — qui imitent le processus d'apprentissage du cerveau et que l'on peut définir comme un processus d'automatisation de la prédiction. Concernant l'HTR, le *Deep Learning* a permis au processus de reconnaissance de s'affiner au niveau de la segmentation du texte, soit la reconnaissance des caractères, des mots, des lignes et des paragraphes[44].

Cela a notamment permis d'améliorer la reconnaissance des documents historiques manuscrits et a donné lieu à la création de deux conférences internationales majeures dans ce domaine ; *International Conference of Document Analysis et Recognition and the International Conference on Frontiers in Handwriting Recognition*[45]. On retrouve également le projet européen *Recognition and Enrichment of Archival Documents* (READ) qui est notamment à l'origine de la plateforme Transkribus qui développe une technologie avancée de reconnaissance de texte sur la base de réseaux neuronaux artificiels et qui aboutit à une infrastructure accessible au public[46]. Si aujourd'hui la plateforme offre effectivement au public la reconnaissance automatique de texte pour des corpus limités, elle a récemment introduit une version payante pour les larges corpus de source, nous poussant à nous tourner vers une alternative gratuite.

L'HTR peut avoir lieu *online* ou *offline*, c'est-à-dire que le premier a lieu lorsque la reconnaissance s'effectue au moment même où la personne écrit sur un support numérique, permettant ainsi l'accès aux informations métriques et temporelles[47]. Ici nous nous intéressons à la reconnaissance automatique de texte *offline*, soit lorsque le texte est manuscrit et ensuite numérisé et que le traitement se fait au niveau de l'image. L'un des enjeux de la reconnaissance automatique de texte touche à la nature cursive de l'écriture manuscrite, à la variété de taille et de formes des caractères utilisés ainsi qu'à l'utilisation de vocabulaires divers[48].

---

43 Purcell Jake, « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », vol. 7, nº 0, Ubiquity Press, 2021, p. 2

44 Sousa Neto Arthur Flor de *et alii*, « Towards the Natural Language Processing as Spelling Correction for Offline Handwritten Text Recognition Systems », *Applied Sciences*, vol. 10, nº 21, Multidisciplinary Digital Publishing Institute, 2020, p. 2

45 Muehlberger Guenter et alii, « Transforming scholarship… », *art. cit.*, p. 956

46 Muehlberger Guenter et alii, « Transforming scholarship… », *art. cit.*, p. 955

47 Scheidl Harald, *Handwritten Text Recognition in Historical Documents*, Vienne : Technische Universität Wien), 2018, 80 p. 1

48 *Ibid.*

Pour rendre possible le *Deep Learning*, il faut fournir à l'algorithme un *training set* qui dans notre cas correspond à un certain nombre de transcriptions qui ont été faites par nos soins des cahiers de Jean Henri Polier de Vernand. Ce *training set* doit être représentatif des différentes parties d'une collection d'archives, reflétant une variété appropriée de mises en page, de vocabulaire et de styles d'écriture et devient ensuite ce que l'on appelle *ground truth* — un terme couramment utilisé en *Machine Learning* pour désigner des informations exactes et objectives fournies par des processus empiriques directs[49].

Ces données d'entraînement constituées des images numérisées et de leurs transcriptions sont utilisées comme données de références afin de constituer un modèle de reconnaissance automatique d'écriture manuscrite qui servira à retranscrire le document historique. Le texte retranscrit est lié aux images grâce à une analyse de mise en page — *Layout Analysis* — qui détecte la segmentation, c'est-à-dire les zones de texte au niveau des blocs de texte, les zones de lignes du texte et les lignes de base du texte[50]. Différents facteurs peuvent influer sur cette partie de l'HTR sur des documents historiques comme le style d'écriture ou les inconsistances de mises en page de la source, l'inclinaison des lignes[51].

Pour évaluer un modèle de reconnaissance automatique d'écriture manuscrite, on fait appel au taux d'erreur de la reconnaissance des caractères et au taux d'erreur sur les mots. Character Error Rate (CER) est définit comme le nombre minimum d'opérations de modification au niveau des caractères qu'un mot doit faire correspondre au *ground truth*. Word Error Rate (WER) est défini de la même manière, mais au niveau du mot dans la phrase[52]. De manière générale, on considère qu'un taux d'erreur de caractères en dessous de 10% peut être exploitable et qu'en dessous de 5% il est très bon et que les erreurs restantes sont généralement dues à des mots rares ou inconnus[53].

---

[49] Muehlberger Guenter *et alii,* « Transforming scholarship… », *art. cit*., p. 959

[50] *Ibid.*

[51] Gatos Basilis *et alii*, « Segmentation of Historical Handwritten Documents into Text Zones and Text Lines », in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Greece : IEEE, 2014, p. 464

[52] Sousa Neto Arthur Flor de et alii, « HTR-Flor++: A Handwritten Text Recognition System… », *art. cit*., p. 3

[53] Purcell Jake, « General Models for Handwritten Text Recognition… », *art. cit.*, p. 2

3. JEAN-HENRI POLIER DE VERNAND (1715-1791)

3.1. INTRODUCTION : LAUSANNE AU XVIIIE SIÈCLE

Sans refaire ici l'histoire du pays de Vaud sous l'Ancien Régime, notons que dès 1535 les Bernois prennent possession de Lausanne et y imposent la Réforme[54]. Dès lors, «il y un nouveau souverain, Leurs Excellences de Berne, et une nouvelle confession, le protestantisme [55]. Cette appellation souvent abrégée LL. EE. était utilisées pour désigner le gouvernement bernois et ses membres et peut être considéré comme équivalent à un titre de [56]. Berne crée ainsi un «État extérieur» où il reproduit le modèle du Gouvernement de Berne et place son contrôle entre les mains de membres des familles patriciennes tout en séparant le Pays de Vaud en différents bailliages[57]. Chacun d'entre eux est dirigé par un bailli — qu'on appelait parfois gouverneur ou préfet — qui agissait en tant que représentant de la République de Berne et qui agissait en son nom[58]. Celui-ci était nommé pour six ans par le conseil des Deux-Cents à Berne et détenait de nombreux pouvoirs, à la fois judiciaires, militaires et fiscaux[59].

À Lausanne, le bailli prenait ses quartiers dans l'ancienne résidence de l'évêque de Lausanne et ce représentant de la République de Berne percevait des revenus importants ainsi que beaucoup de prestige[60]. À ce titre, Lausanne était bien plus modeste en termes de fortune que ne l'était Berne ou Genève par exemple et était «avant tout un centre vinicole, agricole et intellectuel grâce à son Académie et à ses imprimeurs, en même temps qu'aristocratique et un lieu de villégiature pour étrangers»[61].

D'un point de vue géographique, « le bailliage [de Lausanne] de s'étendait de la Venoge à la Veveyse, et des rives du Lac aux forêts du Jorat »[62]. Si dans le Pays de Vaud on retrouvait des seigneuries individuelles détenues par des membres de l'aristocratie et qui pouvait dans une certaine mesure avoir des droits de justice, Lausanne peut être considérée comme une seigneurie collective qui comprenait des villages soumis, des cours de justice et un

---

[54] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 16
[55] Chuard Corinne, *Histoire vaudoise: un survol*, Gollion : Infolio, 2019, p. 63
[56] *Ibid.*, p. 64
[57] *Ibid*., p. 307
[58] *Ibid*.
[59] Biaudet Jean-Charles *et alii*, *Histoire de Lausanne*, Toulouse : Privat, 1982, p. 176
[60] *Ibid*., p. 210
[61] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 31
[62] Biaudet Jean-Charles et alii, *Histoire de Lausanne*, *op. cit*., p. 176

consistoire dont le nom, jadis chapitre, « marque bien la provenance des terres de cette seigneurie, précédée des “largitions” faites par Berne à ses anciens combourgeois »[63].

3.2. GÉNÉALOGIE

La date de naissance précise de Jean-Henri Polier de Vernand n’est pas connue avec certitude, mais celui-ci note qu’il est baptisé le 15 mars 1715 et «on peut donc admettre, sans une grande marge d’erreur, qu’il est né le 4 mars, puisqu’à cette époque on procédait rapidement à cette cérémonie à cause de la grande mortalité infantile»[64]. Il descend d’une famille de noble française et «le premier qui vint en Suisse fut Jean de Polier envoyé en 1553 comme secrétaire d’ambassade du roi de France auprès des Ligues suisses et des Grisons»[65]. Dans le contexte de l’ancien régime, les familles de propriétaires exercent un grand rôle politique[66].

Jean-Henri Polier de Vernand est le fils d’Etienne-Bénigne Polier — membre du Conseil des Soixante et des Deux Cents au sein de la municipalité de Lausanne — et de Françoise de Tavel[67]. Cette famille est liée aux familles seigneuriales de Bottens et de Saint-Germain[68]. Du côté maternel, les Tavel représentent une grande fortune et une famille d’influence au sein des autorités bernoises[69]. Son frère, George-Louis seigneur de Vernand fit une carrière militaire. Il deviendra colonel au sein des gardes suisses au service des États généraux puis en 1776 général major[70]. Les deux frères resteront célibataires et n’auront pas de descendance. Cette branche de la lignée des Polier de Vernand s’éteindra donc avec eux[71].

3.3. FORMATION ET PARCOURS

Malgré le rôle important pour la ville de Lausanne que Jean-Henri Polier de Vernand sera amené à jouer plus tard dans sa vie, on ne possède que peu d’informations sur sa jeunesse et sur son parcours qui le mènera à la fonction de lieutenant baillival à l’âge de trente-neuf

---

63 *Ibid*., p. 176
64 Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 37
65 *Ibid*.
66 Radeff Anne, *Lausanne et ses campagnes au 17e siècle*, Lausanne : Bibliothèque historique vaudoise, 1980, p. 259
67 Favez Valérie, *Etude du « Mémorial universel »…, op. cit.*, p.4
68 *Ibid*.
69 *Ibid*.
70 Delédevant Henri, *Le livre d’or des familles vaudoises: répertoire général des familles possédant un droit de bourgeoisie dans le canton de Vaud …*, Lausanne : Ed. Spes, 1923, p. 326
71 Favez Valérie, *Etude du « Mémorial universel »…*, *op. cit*., p. 4

ans. On sait cependant que celui-ci étudia au sein de l'Académie de Lausanne durant trois ans entre 1727 et 1730[72].

Elle est créée juste après l'Édit de Réformation du 25 décembre 1536 qui marque un tournant dans l'histoire de Vaud et de Lausanne, et plus particulièrement en ce qui concerne l'enseignement supérieur. Pour appuyer la Réforme, les autorités bernoises ont rapidement cherché à établir un système éducatif pour former les pasteurs de langue francophone pour mettre en place l'Église vaudoise réformée[73]. Elles ont fait appel à des humanistes et des pédagogues de renom pour assurer cette formation[74].

Bien que l'Académie ait été créée pour former des pasteurs, les professeurs ne se sont pas limités à l'enseignement théologique et ont également accordé une place importante aux lettres profanes et aux humanités. Elle comprenait l'enseignement de la rhétorique, de la dialectique, des mathématiques et de la physique. Ainsi l'Académie est devenue le centre et le foyer intellectuels de l'Église vaudoise, ce qui permet à Lausanne de passer d'une ville ignorante et superstitieuse durant la période de domination savoyarde à une ville relativement savante et lettrée durant la période bernoise[75].

Si l'on ne sait pas précisément précisément quelle éducation Polier suit par la suite, on peut supposer qu'il possède manifestement une formation juridique qu'en cette période où beaucoup de romands se formaient au Droit dans les cantons alémaniques ou en Allemagne qu'il ait fait de même[76].

Valérie Favez émet l'hypothèse qu'il ait pu travailler auprès de son père qui était alors conseiller au sein du Conseil des Vingt-Quatre[77]. Ce que l'on sait avec certitude, c'est que dès 1747 à l'âge de trente-deux ans il est lui-même élu au Conseil des Deux Cents et qu'entre 1743 et 1749 il accède à la fonction d'assesseur baillival[78]. Cette information vient notamment d'une lettre anonyme envoyée à son frère Georges-Louis dans laquelle l'auteur recommande au lieutenant les descendants d'un autre assesseur baillival issu de la famille Montrond :

---

[72] *Ibid*., p. . 6
[73] Biaudet Jean-Charles *et alii, Histoire de Lausanne, op. cit*., p. 165
[74] Kiener Marc, *Dictionnaire des professeurs de l'Académie de Lausanne (1537-1890)*, Lausanne : Université de Lausanne, 2005, p. 5
[75] *Ibid*., p. 282-283
[76] Favez Valérie, *Etude du « Mémorial universel »…, op. cit*., p. 6
[77] *Ibid*.
[78] *Ibid*.

> […] feu Mr. votre digne frère avait eu l'emploi de lieutenant baillival, duquel il a joui depuis 1754 à 1778, année de la mort de Mr. De Montrond, assesseur baillival son ancien, sur qui il l'emportat à Berne. Cette préférence fut sans doute autant l'effet du mérite distingué de feu Mr. votre frère que du crédit et de l'appui qu'il trouvat dans les parents qu'il avait à Berne qui agirent pour lui. Mais pas moins elle fut un grand malheur et un préjudice infini à son aîné en charge de la Cour Baillival, son ami d'ailleurs, et qui avait acheté cet emploi d'assesseur très cher. Il a laissé une famille, des enfants et des petits-enfants nombreux et totalement dénuée de bien[79].

Cette citation peut également laisser à penser que le père de Jean-Henri Polier aurait potentiellement acheté ce poste pour lui. Si en théorie les mandats politiques sous l'ancien régime étaient ouverts à tous les bourgeois de la ville, ceux-ci restaient conditionnés par la coutume où l'on payait en remerciement aux nominations. Ainsi quelques semaines après sa nomination en tant que lieutenant baillival qui a lieu le 27 avril 1754, Polier écrit dans ses cahiers le 3 mai de cette même année :

> Remis à Mr. Le Blf[80] et à Mme pour le remercier de m'avoir donné auprès de LL. EE. la recommandation à la charge de la charge de lieutenant baillival, 25 doubles louis neufs = 800 L.[81]

Qui plus est, l'influence de ses oncles maternels — qui furent tous à un moment donné de leurs vies en charge d'un bailliage bernois[82] — a très certainement influencé sa nomination au poste de lieutenant baillival. On constate donc que, bien qu'il soit romand, les liens avec les autorités bernoises de sa famille ont leur importance quant au choix que représente Jean-Henri Polier de Vernand.

3.4. POSITION SOCIALE

À Lausanne au XVIIIe siècle, les familles aptes à occuper des postes municipaux ou des fonctions de magistrats trouvaient du travail en exerçant leurs fonctions, bien que cela leur laissât encore des temps libres. Jean-Henri Polier de Vernand était avant tout un riche propriétaire foncier et occupait donc la majorité de son temps à travers de multiples mandats qu'il occupa au fil de sa vie.

---

79 ACV : P Monod 413 cité par MORREN Pierre, *La vie lausannoise au XVIIIe siècle…, op cit*, p. 38-39
80 Bailli, ici Samuel Moutach qui occupa cette fonction entre 1750 et 1755
81 ACV P René Morod 1
82 Favez Valérie, Etude du « Mémorial universel »…, *op. cit*., p. 7

On sait grâce à la thèse de l'historien américain Jeremy Charles Jackson sur la politique municipale lausannoise du XVIe siècle jusqu'à la fin de l'ancien régime que son organisation était fondamentalement oligarchique[83]. Cela signifie notamment que ces derniers ne vivaient pas de leurs fonctions, impliquant ainsi une forme de sanction censitaire à l'accès des mandats politique. Ainsi Polier, bien que lieutenant baillival pendant trente-sept ans, ne tira presque pas de rémunération de ce poste :

> Les fonctions officielles étaient plus ou moins gratuites […]. Les revenus de Jean-Henri Polier de Vernand provenaient de trois sources principales : ceux qu'il retirait de ses multiples fonctions, ceux provenant de son appartenance aux Conseils des Soixante et Deux Cents, enfin, pour la plus grande part, ceux qu'il retirait de ses propriétés.[84]

En effet, dans ce système oligarchique qu'était la municipalité de Lausanne durant l'ancien régime, «cette quasi-gratuité était en quelque sorte la contrepartie de la possession d'une certaine fortune, qu'il fallait prouver, pour pouvoir prétendre à une place dans les conseils de la ville»[85]. Ainsi pour l'année 1756 par exemple il était payé, en plus des redevances en nature représentant une coupe d'avoine et 11 sacs de froments, 92 florins et 6 pour le poste de lieutenant baillival, 387 florins et 6 comme conseiller baillival[86].

Ses revenus personnels venaient principalement de ses propriétés, en plus de «quelques fonds déposés chez deux banquiers déposés à Paris, notamment 3 actions de la Compagnie des Indes dont il retirait 2 à 300 livres annuellement»[87]. Il louait également deux appartements au sein de l'hôtel particulier qu'il habitait à la rue de Bourg en plus de trois petits logis et de deux boutiques et de six autres immeubles dont il partageait la possession avec son frère[88]. Jean-Henri Polier de Vernand étant également le seigneur d'un fief, «la plus importante part de ses rentrées provenait du domaine de Vernand et de ses vignobles»[89].

Finalement, il était également membre de la Société du Cercle de la rue de Bourg dès sa fondation en 1761, conçu comme un cercle de sociabilité typique du XVIIIe siècle[90]. Dans le contexte des Lumières en Suisse romande, «les salons mondains, avec leur esprit

---

[83] ackson Jeremy Charles, *The Evolution of a Municipal Oligarchy: Lausanne, 1536-1798*, Ann Arbor Mich : Univ. Microfilms international, 1977
[84] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op. cit*, p. 63
[85] *Ibid*., p. 63
[86] *Ibid*., p. 64
[87] *Ibid*., p. 65
[88] *Ibid*., p. 68
[89] *Ibid*., p. 65
[90] Charrière W. de, « Le cercle de la rue de Bourg fondé en 1761 », *Société vaudoise d'histoire et d'archéologie*, 1914, DOI: 10.5169/SEALS-19506.

d'hospitalité, de civilité et de divertissement, y contribuent aussi à leur manière»[91]. Il était également membre et président de la Société économique de Lausanne inspirée de sa consœur bernoise[92].

3.5. POSITIONS POLITIQUES

Jean-Henri Polier occupe également des mandats politiques au sein de l'organisation municipale de Lausanne qui se caractérise par sa place particulière durant la période bernoise vis-à-vis du reste du pays de Vaud. Si le bailli est le représentant tout puissant de la République de Berne et que, au moyen de son lieutenant, d'assesseurs, d'un secrétaire et d'un trésorier, se trouve au sommet de l'administration lausannoise, Lausanne maintient certains privilèges et l'organisation municipale[93]. La ville a la compétence juridique de haute, moyenne et basse justice, même si certains verdicts doivent être approuvés par le Sénat de Berne. Cela se fait au travers du Conseil des Vingt-Quatre, des Soixante et des Deux Cents — auxquels, à l'exception du premier, participe Jean-Henri Polier.

Le premier, le Conseil des Vingt-Quatre est le conseil le plus ancien de Lausanne découle d'un accord conclu le 6 juillet 1481 entre le prince-évêque et les bourgeois de la ville basse[94]. Les sources parlent généralement du «Conseil» avec une majuscule et ses membres se font appeler les «conseillers»[95]. Depuis la fin du XVI[e] siècle, le Conseil des Vingt-Quatre a pris le contrôle et a acquis une position dominante sur les autres instances de la ville[96].

Le Conseil des Vingt-Quatre assume la majeure partie des décisions politiques importantes et a un certain contrôle sur les autres conseils. De plus, il a des responsabilités judiciaires importantes, sur lesquelles nous reviendrons plus tard. Certains de ses membres composent la Chambre des Vingt-Quatre, qui entend les appels en première instance pour les affaires civiles. Une autre fonction importante du Conseil est «d'empêcher l'accaparement et fixe les prix de vente du pain, de la viande, du poisson ou du vin, parfois même des oranges, des citrons, du savon ou des chandelles»[97].

---

91 Chuard Corinne, *Histoire vaudoise…, op. cit.,* p. 85
92 Abetel Emmanuel, « Polier de Vernand, Jean-Henri », *art. cit.*
93 MORREN Pierre, *La vie lausannoise au XVIIIe siècle…, op. cit*, p. 16-17
94 *Ibid*., p. p. 17
95 Biaudet Jean-Charles *et alii, Histoire de Lausanne*, *op. cit*., p. 221
96 *Ibid*., p. 221
97 *Ibid*., p. 221

Le deuxième Conseil est celui des Soixante au sein duquel Polier est à la fois représentant du bailli et également comme conseiller nommé par ses pairs. En charge de l'administration et de l'admission de la bourgeoisie à Lausanne, certains de ses membres forment également une chambre d'appel pour les affaires civiles suite à celle de la chambre des Vingt-Quatre et avant celle de Berne[98].

Finalement, le Lausannois est également membre du Conseil des Deux Cents qui se compose des membres du Conseil des Vingt-Quatre et des Soixante :

> Chargé à l'origine de représenter la bourgeoisie, ce dernier n'est plus guère convoqué dans la première moitié du XVIIe siècle. Son rôle à cette époque se limite à la confirmation des décisions prises par le XXIV et le LX. Il reprendra pourtant de l'importance dans les dernières décennies du XVIIe siècle et au XVIIIe siècle.[99]

Les critères pour être membre des conseils à Lausanne sont les suivants : être bourgeois depuis au moins dix ans, posséder un minimum de biens immobiliers et être un homme indépendant — excluant ainsi les domestiques ayant obtenu la bourgeoisie[100].

Au XVIIIe siècle, le Conseil des Deux Cents est responsable de nommer les membres du Conseil des Vingt-Quatre et les hauts fonctionnaires et confirme les charges des membres de tous les conseils chaque année. En revanche, le Conseil des Vingt-Quatre est compétent pour élire de nouveaux membres pour le conseil des Deux Cents et celui des Soixante[101].

Finalement, Polier de Vernand est également membre de la commission de vérification des comptes de l'Hôpital, de la chambre des auditeurs des comptes et de celle des orphelins. Ces chambres sont une particularité lausannoise puisque, à partir du milieu du XVIIe siècle, la municipalité «crée des chambres spécialisées dans des domaines où l'État n'était jamais intervenu précédemment de façon suivie»[102]. Elles sont généralement constituées de membres des trois conseils de l'organisation municipale de la ville. Cependant, la chambre des orphelins — dont la charge est de surveiller les activités des tuteurs et des curateurs — est créée par Berne en 1669 et les autorités bernoises y imposent la présence du lieutenant baillival[103].

---

98 *Ibid*.
99 *Ibid*.
100 Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op. cit*, p. 17-18
101 Biaudet Jean-Charles *et alii, Histoire de Lausanne, op. cit*., p. 221
102 *Ibid*., p. 225
103 *Ibid*., p. 226

3.6. POSITIONS ADMINISTRATIVES

3.6.1. Cours civiles et criminelles

Les fonctions de lieutenant baillival occupées par Polier impliquaient notamment de siéger dans plusieurs cours de justice au sein du bailliage lausannois. Avec la conquête bernoise, la complexité de l'organisation de la justice héritée du Moyen Âge est amplifiée par la coexistence de tribunaux seigneuriaux, ecclésiastiques et civils, auxquels les autorités de Berne ajoutent de nouvelles instances judiciaires. Dès leurs arrivées dans le Pays de Vaud, les autorités bernoises «respectent, dans l'ensemble, les droits et privilèges qui prévalaient sous le régime précédent… tant et aussi longtemps qu'ils ne vont pas à l'encontre du droit bernois»[104]. La juridiction spéciale des évêques et des abbés a disparu avec leurs possessions temporelles et certains nobles deviennent des vassaux de Berne et conservent leur juridiction seigneuriale au sein des différents bailliages qui sont créés en tant qu'unités administratives et judiciaires[105].

L'un des aspects majeurs de l'État durant l'ancien régime et plus particulièrement avec l'arrivée des Lumières est celui de renforcer son pouvoir en contrôlant le système judiciaire de la société, et ainsi garantir une justice plus équitable grâce à ses actions[106]. Cette tâche était cependant complexifiée par la coexistence sous l'ancien régime d'instances baillivales d'un côté et communale de l'autre. Les peines encourues pouvaient aller de l'amende – y compris parfois la via la saisie des biens – à la prison ou des sévices corporels en allant jusqu'à la peine de mort[107].

Concernant les sévices corporels, ils pouvaient être « le carcan, la bastonnade, le pilori, la marque au fer rouge, la mutilation, la potence, le bûcher, la noyade, la roue, l'échafaud»[108]. La torture était régulièrement utilisée tout au long du XVIIIe siècle et considérée comme une procédure permettant d'obtenir un aveu ou encore démontrant l'innocence en cas d'absence de ces derniers, bien qu'il faille l'autorisation spéciale de Berne, qui contrôle de plus en plus en plus son usage[109].

---

104 Chuard Corinne, *Histoire vaudoise…, op. cit.,* p. 72
105 Biaudet Jean-Charles *et alii, Histoire de Lausanne, op. cit*., p. 222
106 Salvi Elisabeth, « La justice de LL. EE. au siècle des Lumières », *op. cit*., p. 333
107 *Ibid.*
108 *Ibid*.
109 *Ibid.*

Au fur et à mesure du XVIIIe siècle, le but de la justice n'était plus simplement de punir, mais également de changer les comportements, notamment en ce qui concerne la violence : le tribunal, quel que soit son type, ne servait pas seulement à régler les différends entre les individus, il était également un lieu important où se jouaient les relations entre le pouvoir en place et les communautés locales[110].

C'est notamment dû à la complexité du système judiciaire qui voit en coexistence différents types de tribunaux allant de cours baillivales, seigneuriales, de châtellenies ou encore des consistoires, et «afin de renforcer la hiérarchie judiciaire, de la cour inférieure à celle du bailliage, l'État devient le principal organe de régulation des conflits»[111]. Berne se voit par exemple obliger d'instaurer une commission permanente pour les affaires criminelles :

> L'ordonnance du 28 août 1704 instaure une commission criminelle permanente chargée de vérifier toutes les condamnations avant leur application par les cours inférieures. Elle renvoie les affaires mal instruites ou manquant de preuves, elle décide de la mise à la torture, vérifie la peine finale, sermonne une cour trop "indolente ou endormie" ou demande la réouverture d'un procès. Telle la justice du roi en France, celle du patriciat bernois prend le contrôle sur l'arbitraire judiciaire. Il n'y a donc pas de recours en matière criminelle, si ce n'est le recours en grâce adressé au bailli, qui transmet la sentence contestée au Petit Conseil.[112]

Qui plus est, le bailli possède également un droit de grâce. Elle consiste en général à réduire une peine en imposant un châtiment moins dégradant ou à raccourcir la durée d'une période de bannissement ou d'incarcération[113]. Les motifs invoqués pour une telle clémence sont généralement liés aux circonstances entourant l'infraction ou à la conduite du prévenu au cours de la procédure judiciaire. Certaines caractéristiques de la personne condamnée, comme l'âge ou l'état civil, peuvent également inciter le souverain à faire preuve de clémence[114].

Lausanne occupait dans ce tableau une position particulière puisqu'elle est parvenue à obtenir des privilèges plus étendus que Payerne par exemple : «ici le nombre des cours s'était multiplié; les unes relevaient de la ville, les autres du gouvernement»[115]. Les tribunaux

---

110 *Ibid*., p. 327
111 *Ibid.*
112 *Ibid*., p. 333
113 *Ibid*.
114 *Ibid*.
115 Maillefer Paul*, Histoire du Canton de Vaud…, op. cit.*, p. 321

baillivaux exerçaient le pouvoir du souverain bernois dans les villages du bailliage qui ne relevaient pas de la ville, ainsi que dans la partie de la ville où LL. EE. avaient réservé la juridiction bernoise[116]. En effet, la ville a «conserver la juridiction sur son territoire (sauf la Cité) et sur un certain nombre de villages de la banlieue»[117].

Ainsi en bas de l'échelle judiciaire en matière civile on retrouve à Lausanne la cour du jadis chapitre où siège Polier de Vernand en tant que représentant du bailli. En effet, «en bas de l'échelle, et relevant de LL. EE., est placé la cour du jadis chapitre, dont la juridiction s'étend sur le quartier de la Cité et sur quelques villages»[118]. Pour ce qui est de la cour civile de la municipalité de Lausanne, elle fonctionnait comme les autres seigneuries du Pays de Vaud — Lausanne représentant à l'époque une forme de seigneurie collective. On retrouve donc également une cour avec un châtelain ou un juge et des justiciers — nommé par le bailli — qui forment la première instance civile[119].

On retrouve ensuite la cour baillivale qui se compose du bailli, de son lieutenant baillival — et donc Polier —, de trois assesseurs et d'un secrétaire[120]. En matière civile, les cours baillivales étaient au-dessus de cours de châtellenies et on en retrouvait une par bailliage. Pour Lausanne, «la seconde instance civile est le tribunal baillival — pour les cours relevant du gouvernement, — la chambre des Vingt-Quatre en appellation — pour les cours relevant de la ville»[121].

Cette chambre était composée de certains des membres du conseil municipal lausannois des Vingt-Quatre qui était la première instance d'appel en matière civile. La seconde était la chambre d'appel des Soixante, ici encore composée de certains membres du conseil du même nom et qui s'occupait donc des deuxièmes appels au niveau civils, la plus haute cour d'appel avant de s'en remettre aux autorités bernoises[122].

Pour les affaires criminelles lausannoises, la désignation du tribunal compétent pour l'audience et le jugement dépendait du bailli ou en cas de conflits de la commission criminelle du Sénat de Berne (petit conseil). Il s'agissait d'une commission permanente créée en 1704

116 Biaudet Jean-Charles *et alii, Histoire de Lausanne*, *op. cit*., p. 222
117 Maillefer Paul*, Histoire du Canton de Vaud…, op. cit.*, p. 321
118 *Ibid*., p. 321
119 *Ibid*., p. 322
120 Biaudet Jean-Charles *et alii, Histoire de Lausanne, op. cit.*, p. 223
121 *Ibid*., p. 322
122 Biaudet Jean-Charles *et alii, Histoire de Lausanne, op. cit*., p. 221

qui était composée de trois puis quatre conseillers bernois[123]. On retrouvait à Lausanne deux cours criminelles en charge de l'instruction selon que le délit ait eu lieu dans la circonscription du bailli ou celle de la ville. Jean-Hier Polier de Vernand faisait partie des deux.

Au sein des instances baillivales, on retrouvait la cour criminelle du château — qui jusqu'en 1730 portait le nom de Cour impériale. Elle était composée du lieutenant baillival, de trois assesseurs baillivaux, de quatre commis de la ville — trois membres du Conseil des Vingt-Quatre et un banneret désigné par ce même conseil — et d'un secrétaire[124].

Au niveau de la municipalité de Lausanne, on retrouvait également la Chambre ou Cour de l'examen des criminels de la ville qui se compose d'un juge et d'un lieutenant de la cour inférieur, d'un secrétaire, de quatre commis (trois conseillers du Vingt-Quatre et un banneret désigné le Conseil et finalement de deux justiciers sans voix délibératives[125]. C'est eux qui sont chargés de l'instruction en matière pénale et il incombe à cette cour «la connaissance définitive de certaines infractions de moindre poids»[126]

3.6.2. Tribunal de la rue de Bourg

En tant que propriétaire de la rue de Bourg, Jean-Henri Polier de Vernand siégeait également au tribunal du même nom, particularité lausannoise sous l'ancien régime. Si l'instruction des enquêtes criminelles relevait soit des instances baillivales — la cour criminelle du château — ou des instances municipales — la Chambre d'examen des criminels de la ville —, «le prononcé du jugement appartenait aux jurés de Bourg, convoqué au château dans le premier cas, à l'hôtel de ville dans le second»[127]. La particularité de ce tribunal durant cette période de domination bernoise tient notamment au fait que si dans le premier cas (cour baillivale) les sentences sont envoyées à Berne pour y être sanctionnées, dans le second le Tribunal de la Rue de Bourg jugeait en dernière instance — bien que le bailli garde son droit de grâce128.

Cette exception de la juridiction baillivale fait dire à l'écrivain suisse Jean-Rodolphe Sinner de Ballaigues en 1787 :

---

123 *Ibid*., p. 223
124 *Ibid*.
125 *Ibid*.
126 *Ibid*., p. 221
127 Maillefer Paul, *Histoire du Canton de Vaud*…, *op. cit.*, p. 322
128 *Ibid*.

> Sa juridiction ne s'étend pas sur la ville et sa banlieue. Le droit attaché aux habitants d'une seule rue, qu'on appelle «Bourg», de juger en dernier ressort les causes criminelles et les délits capitaux qui se commettent dans l'enceinte de la ville est remarquable. Chaque chef de maison a le droit de suffrage dans ce tribunal qui s'assemble publiquement dans la rue.[129]

Le tribunal est composé des bourgeois et des citoyens (fils de bourgeois) propriétaires de la rue de Bourg et «en compensation de cette charge de juré, ces maisons étaient exemptées de droit de mutation (lauds)»[130]. Ce tribunal était assisté par le Conseil des Vingt-Quatre et par un juge[131].

Si le Tribunal de la Rue de Bourg peut en théorie juger sans recours à Berne les affaires qui touchent à sa circonscription, on peut noter que «les affaires graves lui échappent de plus en plus au profit de la cour criminelle du château placée sous le contrôle du bailli»[132].

3.6.3. Consistoire

En tant que lieutenant baillival, Polier de Vernand s'occupait également t de remplacer le bailli à la présidente du vénérable consistoire, place qui lui sera laissée par la majorité des baillis que le lausannois servira[133]. L'introduction des consistoires par les autorités bernoises était notamment l'un des instruments utilisés pour instaurer la réforme dans le pays vaudois :

> Les consistoires ont en effet été pensés dès leur création, avec le concours des théologiens et des pasteurs, comme un instrument disciplinaire chargé d'inculquer aux populations des normes religieuses et morales.[134]

Les consistoires locaux étaient placés sous la direction du consistoire supérieur de Berne qui «veillait avec eux au maintien des bonnes mœurs, de la doctrine chrétienne officielle»[135]. Cette juridiction ecclésiastique se chargeait des mariages, des divorces, des enfants naturels et des questions de mœurs[136]. Ici encore, Lausanne occupe une position différente du reste du pays de Vaud :

---

[129] Sinner de Ballaigues, *Voyage historique et littéraire en Suisse occidentale, Tomme II,* 1787, cité par Spalinger René, *Quand Mozart passait à Lausanne: chronique inédite*, Genève : Slatkine, 2006, p. 84-85
[130] Biaudet Jean-Charles *et alii, Histoire de Lausanne, op. cit*., p. 223
[131] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op. cit*, p. 17
[132] Salvi Elisabeth, « La justice de LL. EE. au siècle des Lumières », *op. cit*., p. 331
[133] Favez Valérie, *Etude du « Mémorial universel »*…, *op. cit*., p. 5
[134] Tosato-Rigo Daniele et Staremberg Goy Nicole, « Avant-propos », in Tosato-Rigo Daniele et Staremberg Goy Nicole, *Sous l'oeil du consistoire. Sources consistoriales et histoire du contrôle social sous l'Ancien Régime*, Etudes de Lettres, 2004, p. 5
[135] *Ibid*., p. 305
[136] *Ibid*., p. 323

> Mais ce qui était particulier à Lausanne, c'est que le tribunal baillival, qui fonctionnait par ailleurs que comme tribunal d'appel des cours de justice inférieures, était en outre un tribunal de première instance pour une partie de la ville, qui était ainsi détachée et soustraite aux autorités municipales; à savoir la Cité proprement dite, le quartier qui s'étend du château (résidence autrefois des évêques et maintenant du bailli) à l'esplanade de la cathédrale, au haut des Escaliers-du-Marché. La Cité était sous l'administration directe du bailli, avec ses habitants. C'est ce qui explique notamment l'existence d'un consistoire baillival à côté du consistoire de la ville de Lausanne.[137]

Dans cette configuration, la présidence du consistoire revenait alors au bailli, qui le plus souvent confiait cette charge à son lieutenant baillival, «son adjoint, et remplaçant choisi parmi les familles nobles du lieu, qui siège en raison de sa charge dans les principales instances communales»[138]. Durant la majorité du mandat de Polier, cette place de président du Consistoire lui était laissée par le bailli en place.

[137] Biaudet Jean-Charles *et alii, Histoire de Lausanne*, *op. cit*., p. 176

[138] Staremberg Goy Nicole, « Contenir la parole et le geste à Lausanne au XVIIIe siècle. Le Consistoire de la Ville face à la violence », in Tosato-Rigo Daniele et S Staremberg Goy Nicole, *Sous l'oeil du consistoire…, op. cit*, p. 179

4. LES «CAHIERS POLIER»

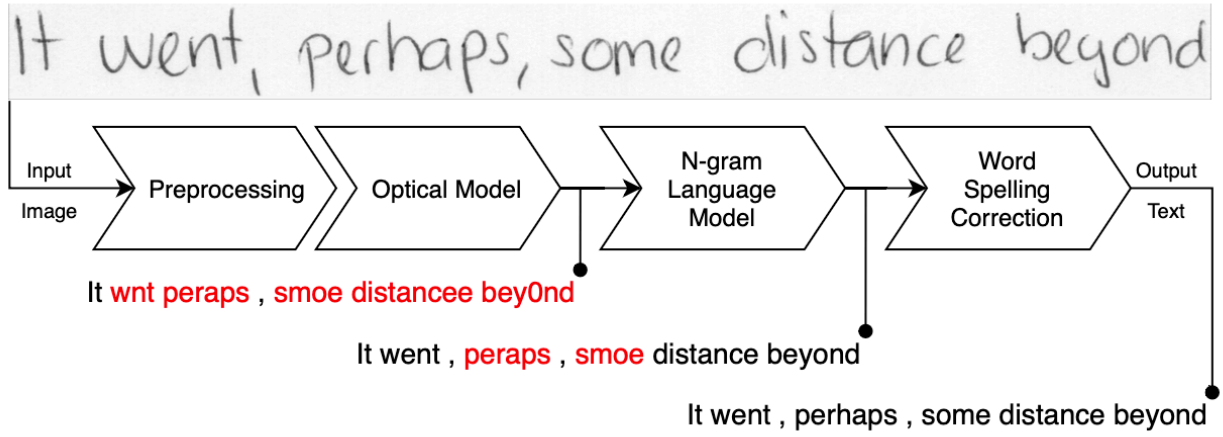
4.1. MÉTHODOLOGIE SUIVIE

4.1.1. HTR-Flor++

Dans ce travail, nous nous sommes appuyés sur le modèle de reconnaissance automatique de texte HTR-Flor++ développé par Arthur Flor de Sousa Neto. Celui-ci présente l'avantage de ne pas être basé sur le modèle des Markov cachés, mais sur le *Convolutional Recurrent Neural Networks* (CRNN). Ce modèle rend le processus de reconnaissance plus efficace au niveau de la segmentation[139]. De plus, il s'agit d'un système de reconnaissance de textes manuscrits basé sur un ensemble de modèles optiques (OCR) et linguistiques :

> In this way, we propose a Convolutional Recurrent Neural Network (Gated-CRNN) architecture combined with two steps of Language Models: (i) the traditional N-gram to correct the text at the character level; then (ii) spelling correction with edit distance based on a word-frequency dictionary. [140]

FIGURE 1



Le système HTR-Flor++

Au moyen de la figure 1, on comprend alors qu'au-delà de la méthode optique de reconnaissance de caractères — en utilisant les images comme input et du texte reconnu

[139] Sousa Neto Arthur Flor de *et alii*, « Towards the Natural Language Processing… », *art. cit.,* p. 2
[140] Sousa Neto Arthur Flor de *et alii,* « HTR-Flor++: A Handwritten Text Recognition System… », *art. cit*., p. 1

comme output[141] —, un modèle linguistique a également été appliqué pour améliorer le résultat final. Premièrement, le modèle de langage *N-gram* permet de minimiser le taux d'erreur par mot et a été entraîné avec les transcriptions faites auparavant. Deuxièmement, un dictionnaire de fréquence de mot — effectuée avec le corpus d'entraînement — est utilisé pour corriger l'orthographe[142].

L'usage combiné du CRNN et des progrès faits en linguistique dans le champ du traitement automatique des langues — Natural Language Processing (NLP) — permet un meilleur traitement de la reconnaissance automatique d'écriture manuscrite[143].

4.2.2. Le modèle des cahiers Polier

Pour élaborer un modèle de reconnaissance automatique d'écriture manuscrite, il nous a fallu faire un certain nombre de retranscriptions des cahiers de Polier de Vernand pour élaborer des données servant à l'entraînement du modèle. Pour ce faire, nous avons utilisé Transkribus puisque celui-ci permet gratuitement d'effectuer des transcriptions et de faire une analyse de la structure du document — *layout analysis* en anglais — pour détecter la segmentation du texte. Dans ce domaine, de récentes avancées ont permis d'améliorer la précision de ce processus crucial plus adapté à des documents historiques avec une structure complexe[144].

Il a donc fallu retranscrire un certain nombre de pages afin d'obtenir des données d'entraînement représentatives du corpus. En effet, il faut que ce set de données reflète une variété de mises en page, de vocabulaire et de styles d'écriture en sélectionnant soit des pages spécifiques, soit en choisissant des pages à un intervalle régulier[145]. Ainsi, pour élaborer notre set de données pour entraîner notre modèle, nous avons effectué une sélection à la fois pour des pages spécifiques de certains cahiers présentant des particularités ou encore la présence de nombreux chiffres par exemple, mais aussi certaines pages suivant un intervalle de 10 pages. Dès lors, nous avons retranscrit les pages comme indiqué dans le tableau 1 présenté à la page suivante.

---

141 Sousa Neto Arthur Flor de *et alii*, « Towards the Natural Language Processing… », *art. cit.,* p. 3
142 Sousa Neto Arthur Flor de *et alii*, « HTR-Flor++: A Handwritten Text Recognition System… », *art. cit*., p. 2
143 Sousa Neto Arthur Flor de *et alii*, « HTR-Flor: A Deep Learning System for Offline Handwritten Text Recognition », in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, p. 55
144 Muehlberger Guenter *et ali*i, « Transforming scholarship… », *art. cit*., p. 959
145 *Ibid.*

TABLEAU 1

Numéro de cahier :	**Page(s) :**	**Numéro de cahier :**	**Page(s) :**
001	4 – 5 – 6 – 49	090	10
010	10 - 20	100	10 – 30 – 50
020	4 – 5 – 6 – 7 – 8	110	8
040	7 – 8 – 41 – 51	125	50
050	10	145	20 – 30
060	10 – 20 – 30 – 40	155	10 – 20
070	10	160	10 – 19 – 28 – 41
080	10 – 30 – 50	185	10 – 20
Total : 40 pages numérisées retranscrites (double page manuscrites de cahier)			

Afin de créer ce modèle spécifique aux cahiers de Jean-Henri Polier de Vernand, nous avons dans un premier temps utilisé le logiciel Transkribus. Permettant de créer des données d'entraînements et de l'utiliser pour créer un modèle de prédiction, ce logiciel semblait au début de ce travail s'y prêter parfaitement. Cependant, Transkribus a récemment mis en place une limite des pages pouvant être retranscrite gratuitement qui ne nous permettait pas de couvrir les 26 300 pages du Lausannois dans un budget envisageable. En effet, le logiciel offre actuellement la possibilité d'appliquer un modèle HTR sur 500 pages manuscrites gratuitement, puis propose un forfait commençant à 18€ pour 120 crédits et de 2 160€ pour 10 000 crédits — 1 crédit correspondant à 1 page manuscrite.

C'est donc pour cette raison que nous nous sommes tournés vers le modèle de reconnaissance de texte HTR-Flor++ qui, bien que moins facile d'utilisation pour un utilisateur moins familiarisé avec le langage informatique — l'un des avantages principaux de Transkribus —, nous a permis d'obtenir des résultats similaires. En effet, le meilleur test obtenu sur Transkribus nous permettait d'obtenir un modèle avec un CER de 9,57% tandis

que les tests effectués sur HTR-Flor++ arrivent à un CER de 8,78%. Comme nous l'avons vu précédemment, un CER en dessous de 10% est exploitable.

Un CER plus bas n'a pas été possible au vu de la diversité de la qualité d'écriture contenue dans les cahiers Polier, où par exemple un double «n» peut facilement être confondu avec la lettre «u» (figure 2) ou entre la confusion entre un «T» et un «J» (figure 3). Qui plus est, les cahiers de Polier de Vernand présente également des caractéristiques inhérentes à l'écriture manuscrite qui peuvent être influencée par la fatigue ou le temps à disposition pour écriture qui rende difficile la reconnaissance automatique comme par exemple la lettre «S» en majuscule qui déborde sur la ligne inférieure et peut être interprétée par le programme d'apprentissage comme une lettre supplémentaire au mot adjacent ou encore un «O» mal fermé interprété comme un «C» donnant ainsi «Crphelin» au lieu de «Orphelin».

FIGURE 2

Exemple d'application du modèle sur un cahier non retranscrit

FIGURE 3

Exemple de retranscription après l'apprentissage sur les données d'entraînement

Pour passer des retranscriptions effectuées sur Transkribus à notre modèle d'interprétation d'écriture manuscrite, nous avons exporté de la plateforme la segmentation — soit les différentes lignes détectées comme images au format «JPG» et les retranscriptions comme

texte lié à ces images au format XML. Nous avons ensuite utilisé un script python (Annexe 1) qui permet de charger l'image et d'en extraire le nom et d'obtenir le numéro du carnet et le numéro de page de la source puis de lire le fichier XML généré par Transkribus pour cette page. Il revient ensuite sur chaque région de texte détectée sur chaque page grâce à la *layout analysis* et pour chacune d'entre elles il obtient les coordonnées du contour de la région de texte et de la ligne de base associée et augmente légèrement la marge des pixels autour de la ligne pour ensuite la redresser en suivant la ligne de base — Polier écrivant parfois en biais sur son cahier. Chaque segment est ensuite associé à la transcription faite sur Transkribus.

Finalement, il faut créer des partitions pour les ensembles d'entraînement, de validation et de tests à partir d'une liste d'éléments. L'ensemble d'apprentissages est utilisé pour former le modèle, l'ensemble de validation est utilisé pour évaluer les performances du modèle et le réglage des paramètres pendant l'apprentissage, et l'ensemble de tests est utilisé pour évaluer les performances finales du modèle une fois l'apprentissage et les réglages terminés[146]. Dans ce cas, la proportion est de 10% pour la validation et de 20% pour le test, ce qui signifie que les 70% restants seront utilisés pour l'entraînement.

La reconnaissance automatique de texte se fait ensuite sur le script proposé par Arthur Flor de Sousa Neto[147] qui permet donc d'entraîner un modèle de reconnaissance de texte manuscrit en utilisant TensorFlow — une bibliothèque open source d'apprentissage automatique et de traitement de données, créé par Google — sur un périphérique GPU qui nécessite l'utilisation de Google Colab qui est une plateforme en ligne proposée par l'entreprise du même nom permettant d'effectuer des calculs élevés pour les tâches d'apprentissage automatique (machine learning).

Il définit les paramètres de l'apprentissage du modèle tel que la source de données, l'architecture *flor* — soit le modèle neuronal ou la façon dont les différentes couches de neurones sont organisées et connectées, le nombre d'époques — *epochs* en anglais, soit le nombre d'itération complète de l'ensemble des données d'entraînement, la taille des *batchs*, soit le nombre d'exemples utilisés avant d'ajuster le modèle. Le script divise ensuite les

[146] Sousa Neto Arthur Flor de et alii, « HTR-Flor++: A Handwritten Text Recognition System… », *art. cit*., p. 3
[147] Arthur Flor de Sousa Neto, arthurflor23/handwritten-text-recognition, 08.03.2023 [14 avril 2019], Python, MIT License, https://github.com/arthurflor23/handwritten-text-recognition.

données produites par la transcription faites sur Transkribus en trois sous-ensembles ; l'ensemble d'apprentissages, l'ensemble de validation et l'ensemble de tests.

L'utilisation de ce script se faire en deux partie : une pour l'entraînement du modèle sur les pages retranscrites et l'une pour l'inférence pour le reste des cahiers. Pour l'élaboration du modèle, nous avons également utilisé des données issues du projet tranScriptorium Bentham. Cet ensemble de transcriptions de la collection des manuscrits Bentham écrits par le philosophe du même nom ont été crées dans le cadre d'un concours organisé lors de la conférence ICFHR 2014 pour évaluer la reconnaissance automatique de textes manuscrits. Cet événement à permis d'élaborer un set de donnée dont le taux d'erreur de transcription est passé de 15,0% à 8,6%[148].

Dans notre cas des cahiers de Jean Henri Polier de Vernand, l'utilisation des données de transcriptions de la collection Bentham a non seulement permis d'obtenir un CER plus bas, il est également utile puisque le lieutenant baillival écrivait parfois en anglais, que ce soit dans le cas où il rédigeait des notes de lectures sur des journaux anglais, ou lorsqu'il ne souhaitait pas pouvoir être lu par ses domestiques. Durant cette partie d'entraînement du modèle, le script prend en compte l'ensemble des données connues pour apprendre à reconnaître les motifs de caractères et mettre en place un dictionnaire de mots et évalue le modèle en comparant la transcription fournie aux valeurs détectées par le modèle.

Une fois le modèle entraîner, intervient la phase d'inférence de la reconnaissance automatique de texte. Pour ce faire, nous avons dû dans un premier temps assembler l'intégralité des cahiers de Polier de Vernand dans un format reconnaissable pour le script HTR–Flor++. Le programme d'Arthur Flor de Sousa Neto permet de convertir les données issues de Transkribus en fichiers hdf5 auxquels l'on peut ensuite appliquer le modèle entraîné. Le modèle est alors utilisé pour prédire les segments créés par la *layout analysis* faite sur Transkribus et donc détecter les symboles, que ce soient des lettres ou de la ponctuation. Ainsi, l'intégralité des cahiers a pu être prédite grâce au modèle entraîné sur les transcriptions effectuées par nos soins ainsi que le modèle bentham.

Les données prédites sont alors fournies dans un fichier texte par cahier qu'il convient ensuite de remettre dans l'ordre des détections effectuées par Transkribus. Pour ce faire,

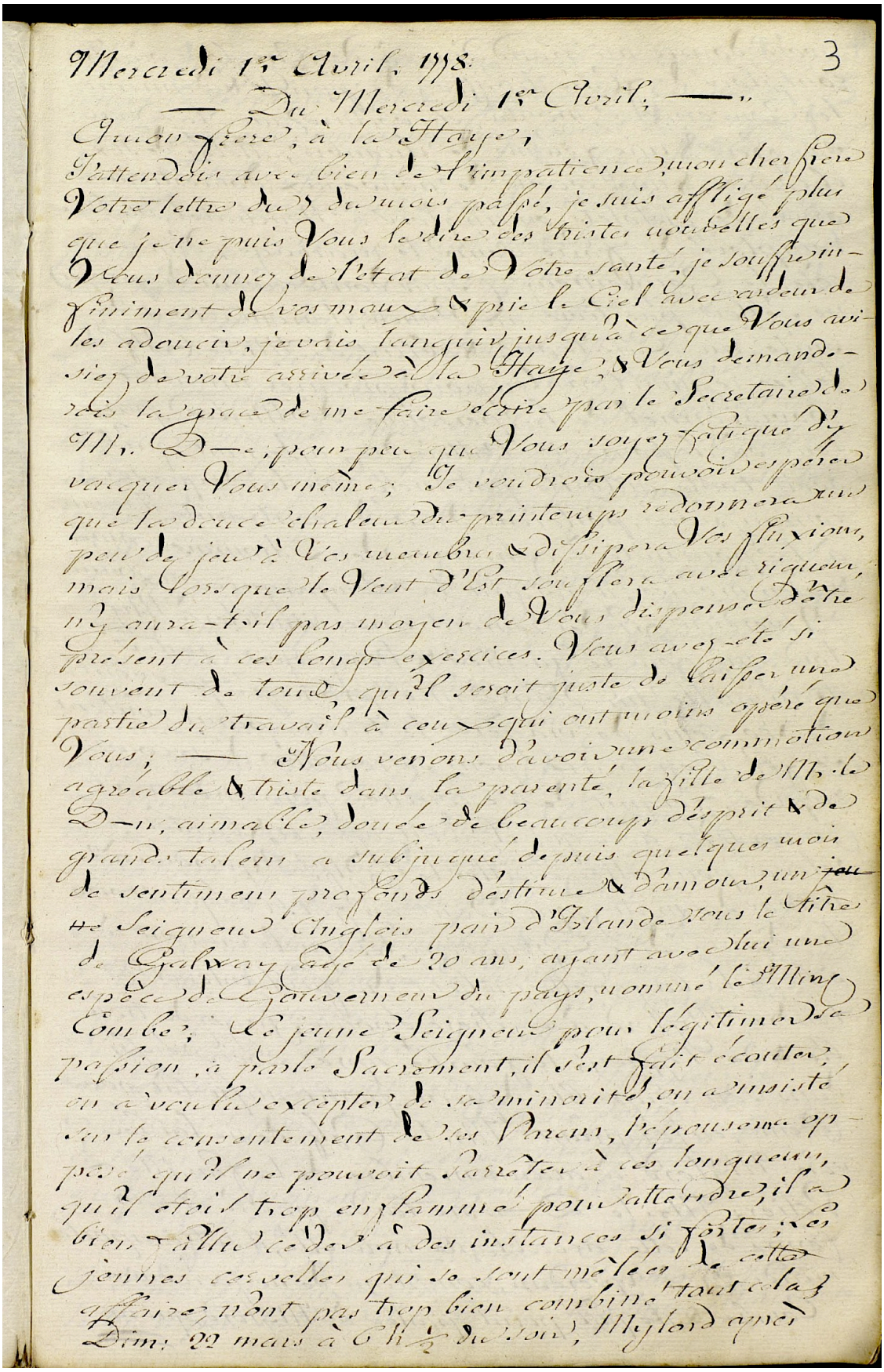
---

[148] Sanchez Joan Andreu *et alii*, « ICFHR2014 Competition on Handwritten Text Recognition on Transcriptorium Datasets (HTRtS) », in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Greece : IEEE, 2014, p. 785-786

nous avons utilisé le script Python présenté en annexe II. Il nous permet de créer des fichiers JSON contenant les prédictions des cahiers en texte continu tout en gardant l'information du numéro du cahier et du numéro de la page de la transcription. L'ordre est reconstitué grâce au fichier créé lors de l'utilisation du script présenté à l'annexe I et donc de l'analyse effectuée par Transkribus. Ici nous avons un exemple du résultat du modèle de reconnaissance automatique d'écriture manuscrite créée pour les cahiers Polier :


```
[
  {
    "cahier_n": 91,
     "page_n": 3,
     "transcription": "mercredi 1 avril ; 1778 \n
    du mercredi 1er avril ;, \n mon frere a la haye; \n
    jattendois aver bien de l'impatience mon cher frere \n votre
    lettre du 7 du mois passe, je suis afflige plus \n que je ne
    puis vous le dine des tristes nouvelles que \n lus donnez de
    l'etat de votre sante, je sousse in \n finiment de ves maux
    & prie le ciel avec ardeur de \n les adoucir, je vais
    lanquir jusqu'a ce que vous avi \n siez de votre arrivee a
    la haye. vous demand. \n rais la place de me faire ecrire
    par le secretaire de \n mr de, pour peu qu vous soyez
    fatique d't \n vaquer vous meme ; je voudrois pouvoir
    esperer \n que la douce chaleur du printemps donnera un \n
    peu de jeu a vos membres & dissipera vos fluxions \n, mais
    lorsque le vent d'est soufler a avec riguenr \n n'y aura
    t'il pas moyen de vous dispenser d'etre \n present a ces
    longr exercices vous avez ete si \n souvent de tour qu'il
    seroit juste de laisser une \n partie du travail a ceux qui
    ont moins apereque \n vous ; nous venons d'avoir une
    commolion \n agreable & triste dans la parente; la fille de
    mr. le \n don, aimable, douee de beaucoup desprit & de \n
    pande tulenr a suljuque depuis quelques mois \n de sentimens
    profonds destime & d'amour un jeu \n n: seigneur anglais
    pair d'irlande sous le titre \n de gatxay, age de 20 ans,
    ayant avec lui mne \n espece de gouverneur du pays, momme le
    minf \n combe; le jeune seigneur pour legitimer se \n
    passion ; parl. sacrement il s'est fait ecouter \n en a
    voulu excopter de sa minorite on a miste \n sur le
    consentement de les parent, l'epouse ma op \n pes qu'il ne
    pouvoit sarreter a ces lonqueurs \n qu'il etois trop
    emflamme pour attendre,; il a \n bien d illu ceder a des
    instances si fortes ; les \n cennes ces velles qui se sont
    melees de cette \n faure n'ont pas trop bien combine taut
    ola \n dim: 2 mars a 6 h 1/2 du soir, mylard apres \n "
  }
]
```

FIGURE 4



Mercredi 1re Avril. 1778. 3

— Du Mercredi 1re Avril. —

Quon père, à la Haye,

J'attendois avec bien de l'impatience quon cher frere Votre lettre du 25 du mois passé, je suis affligé plus que je ne puis Vous le dire des tristes nouvelles que Vous donnez de l'état de Votre santé, je souffre infiniment de Vos maux & prie le Ciel avec ardeur de les adoucir, je vais languir jusqu'à ce que Vous avisiez de votre arrivée à la Haye; & Vous demandez la grace de me faire écrire par le Secretaire de Mr. D—e, pour peu que Vous soyez fatigué d'y vaquer Vous même; Je voudrois pouvoir espérer que la douce chaleur du printemps redonnera aux peu de jeu à Vos membres & dissipera Vos flux[ions], mais lorsque le Vent d'Est soufflera avec rigueur, n'y aura-t-il pas moyen de Vous dispenser d'être présent à ces longs exercices. Vous avez été si souvent de tout qu'il seroit juste de laisser une partie du travail à ceux qui ont moins opéré que Vous; — Nous venons d'avoir une commotion agréable & triste dans la parenté, la fille de Mr. le D—n, aimable, douée de beaucoup d'esprit & de grands talens a subjugué depuis quelques mois de sentimens profonds d'estime & d'amour un jeune Seigneur Anglois pair d'Islande sous le titre de Galway âgé de 20 ans, ayant avec lui une espèce de Gouverneur du pays, nommé le Mlord Combe; Le jeune Seigneur pour légitimer sa passion, a parlé Sacrement, il s'est fait écouter, on a voulu l'excepter de sa minorité, on a insisté sur le consentement de ses Parens, l'épreuve a apparemment qu'il ne pouvoit s'arrêter à ces longueurs, qu'il étoit trop enflammé pour attendre, il a bien fallu céder à des instances si pressantes; les jeunes cervelles qui se sont mêlées de cette affaire n'ont pas trop bien combiné tout cela Dimanche 29 mars à 6 h ½ du soir, Mylord agréé

Extrait de la page 3 du cahier n°93

En comparant la transcription obtenue via le modèle HTR et l'extrait de la source physique des cahiers Polier, on sent rend ainsi compte que même si la transcription n'est pas parfaite, elle reste compréhensible et reconnaissable. Aux données qui ont été prédites par l'algorithme, il a également fallu rajouter une étape de nettoyage.

4.2. NETTOYAGE DES DONNÉES

Étant donnée l'importance de notre corpus de source comprenant environ 26 300 pages manuscrites, l'étape du nettoyage des données est aussi importante que coûteuse en temps. En effet, bien que le modèle soit parvenu à un taux d'erreur acceptable avec un CER de 8,78%, cela implique également un certain nombre d'erreurs qui peuvent peser lourd sur la robustesse des résultats d'analyse. En effet, les modèles de reconnaissance automatique de texte permettent notamment de produire rapidement de grandes quantités de données, mais celles-ci peuvent s'avérer incomplètes, inexactes ou incohérentes et il reste nécessaire d'effectuer un travail de relecture après l'utilisation du *machine learning* afin d'identifier et de corriger un maximum d'erreurs. Cela permet notamment de s'assurer que les données produites soient les plus précises et fiables possibles.

Qui plus est, le but de ce travail est non seulement de comprendre ce que contiennent les «cahiers Polier», mais également de rendre disponibles ces données pour d'autres chercheurs. Dès lors, les données doivent être au plus proche possible de la source originelle et nécessite de la transparence sur le processus de leur constitution et du nettoyage des données. À ce stade du travail, nous avons donc dupliqué les résultats obtenus lors des étapes précédentes en deux pour établir deux sets de données, un premier que l'on peut qualifier de diplomatique, respectant la reproduction dactylographique des documents au plus proches possibles de la source et un deuxième visant à être utilisé pour une analyse secondaire aux moyens d'outils automatisés.

Si le nettoyage des données est essentiel, il n'est pas sans poser de problèmes. En effet, notre corpus étant d'une relative importance, il n'est pas aisé de nettoyer ces données d'autant que certaines inexactitudes nécessitent une correction manuelle coûteuse en temps. Pour ce faire, nous nous sommes donc aidés d'outil automatisé afin d'atténuer les différentes difficultés que représente le nettoyage des données. Nous avons dans un premier temps supprimé tous les accents et mis tout le texte en minuscule afin de faciliter les corrections sans trop trahir la nature du texte d'origine.

Concernant les transcriptions diplomatiques, nous avons procédé en identifiant les entités nommées détectées dans le texte au moyen du script Python présenté en annexe III. Celui-ci extrait la transcription de chaque fichier JSON correspondant à un cahier et l'analyse en utilisant un modèle de traitement du langage naturel en français issu de la bibliothèque `Spacy`. Les entités sont ensuite filtrées pour ne conserver que celles qui apparaissent au moins trois fois, ce qui représente plus ou moins un seuil statistique de représentativité. Sur la base de ces entités identifiées, de nombreuses corrections ont été faites, comme par exemple en remplacer toutes les occurrences «crphelin» par «orphelin». En effet, comme évoqué plus haut, l'un des obstacles majeurs de la reconnaissance automatique d'écriture manuscrite réside dans l'inhérente variabilité de l'écriture qui, pour reprendre cet exemple, rend difficile pour un algorithme la distinction pour un «o» ne formant pas un cercle parfait et un «c». Nous avons ainsi procédé à de multiples corrections manuelles, dans la mesure du temps à notre disposition. Ainsi les transcriptions des cahiers Polier ne sont pas parfaites, mais compréhensibles.

Concernant les transcriptions créées pour l'analyse automatisée, l'uniformisation des données a été plus importante. En effet, un nettoyage qui trahirait légèrement la source d'origine est nécessaire pour les techniques de modélisation thématique telle que l'allocation de dirichlet latent (LDA) sur laquelle nous reviendrons dans l'un des sous-chapitres suivants. En tant que modèle statistique, il dépend fortement de la qualité des données d'entrée. Il est donc nécessaire de prétraiter les données et de les nettoyer avant d'appliquer la méthode LDA. Cela nécessite notamment d'uniformiser certaines abréviations utilisées par Jean Henri Polier de Vernand. Par exemple, dans nos données utilisées pour l'analyse les occurrences telles que «therm» ont été remplacées par «thermometre» et «consist» par «consistoire» afin qu'ils soient compris par les scripts d'analyse comme un seul et même mot. Cela implique également de supprimer la ponctuation et les symboles.

Cette étape est d'autant plus importante que la méthode LDA nécessite une importante puissance de calcul et est coûteuse en temps, ce qui renforce l'importance d'avoir des données nettoyées avant d'utiliser celle-ci. Cela permet de s'assurer que les résultats de la modélisation des données sont précis et significatifs afin d'obtenir des informations sur les sujets et les thèmes présents dans notre corpus de source.

Qui plus est, à la différence des transcriptions diplomatique, ce corpus de données présente l'avantage de ne pas avoir à être au plus proche de la source et donc de ne pas

garder l'information des différentes lignes des cahiers Polier. Nous avons donc pu supprimer les retours à la ligne, enregistré dans les fichiers json sous la forme d'un «\n» et ainsi rassembler les mots séparés sur plusieurs lignes par un tiret.

Une étape supplémentaire a été la mise en place d'une liste de *stopwords,* soit une liste de mots à ne pas prendre en compte dans l'analyse. Ceux-ci sont fréquents dans l'utilisation d'algorithme de traitement du langage naturel, et implique par exemple l'exclusion des articles ou des propositions. En plus des *stopwords* typique du français, il nous a fallu en mettre en place d'autres spécifiques cette fois encore aux cahiers Polier. Cela est notamment dû au fait que la détection des entités sur notre corpus de source n'a pas permis une détection correcte des dates. En effet, si les modèles de traitements de données sont généralement entraînés sur de grands corpus de données, ceux existant en langue française peuvent s'avérer moins performants que ceux en anglais. Ainsi nous avons dû intégrer dans les mots à exclure de l'analyse les jours de la semaine et les mois.

4.2. L'ANALYSE DES ENTITÉS NOMMÉES

L'une des premières entrées possibles au moyen de processus automatisé dans un corpus de donnée tel que celui des 26 300 pages des cahiers Polier est celui de la reconnaissance automatique des entités nommées. Par reconnaissance automatique d'entité nommée nous entendons qui se sont développés comme système de compréhension automatique des documents via la reconnaissance des «éléments informationnels pertinents »[149]. Ainsi le script présenté en annexe IV permet d'extraire du corpus de source les entités nommées (Named Entity Recognition ou NER en anglais) et de les classer dans différentes catégories telles que les personnes, les lieux, les organisations et les entités qui répondent de la catégorie « divers ». Le modèle de traitement automatique du langage naturel utilisé est celui de `spaCy`, une bibliothèque Python open source qui permet d'effectuer des tâches de traitement du langage naturel, notamment en français[150].

Le script définit ensuite deux listes de mots à ne pas prendre en compte dans l'analyse. La première liste contient des mots vides courant en Français, tels que les articles, les conjonctions et les prépositions. La seconde liste, appelée « `polier_stopwords` »,

[149] Nouvel Damien et alii, *Les entités nommées pour le traitement automatique des langues*, ISTE Group, 2015, p. 14

[150] Othmen Dhifallah, « Extraction et entraînement des entités nommées avec spaCy », *Medium*, 18.12.2019, https://medium.com/extraction-et-entrainement-des-entites-nommees-avec-spacy, consulté le 08.04.2023

contient des mots spécifiques qui ne sont pas pertinents pour l'analyse, telle que les mois, et les jours de la semaine. S'il est possible d'entraîner le modèle de reconnaissance des entités nommées pour qu'il reconnaisse les jours et les mois — y compris avec leur orthographe usuelle du XVIIIe telle que 7bre pour septembre ou Xbre pour décembre — car il nous a semblé peu pertinent pour la portée possible d'une analyse de fréquence. Finalement, nous avons également exclu les différentes appellations telles que « monsieur » ou son abréviation « mr » qui était fortement présent dans le corpus et qui nous intéressait peu pour la catégorie « personne » par exemple étant donné qu'ils étaient détectés en dehors du nom de la personne à laquelle elles étaient initialement rattachées dans les transcriptions.

La possibilité d'extraire les entités nommées du texte vient du modèle « `fr_core_news_lg` » issu d'une bibliothèque `spaCy` qui est entraîné pour comprendre la langue française et en extraire des entités nommées. Ici nous en avons retenu quatre, les personnes, les lieux, les organisations et celles trop diverses pour appartenir aux trois autres. Le script utilise ensuite des instructions conditionnelles pour vérifier l'étiquette de chaque entité nommée et l'ajouter à la liste appropriée. Finalement, nous avons également rajouté un filtre pour obtenir de ce script IV les dix entités les plus fréquentes pour avoir un point d'entrée dans le contenu des 26'3000 pages des cahiers Polier.

Comme nous pouvons le voir dans le tableau 2 présenté à la page suivante, le fait que Mr. le bailli soit la personnalité la plus fréquemment mentionnée montre son importance dans le corpus et à quel point les cahiers de Jean Henri Polier de Vernand sont intimement liés à sa fonction de lieutenant baillival. La seconde entité la plus fréquente, « Francois », est plus difficile à expliquer. En plus du fait que le prénom est relativement courant, il semblerait qu'ici il s'agisse à la fois d'un défaut dans la reconnaissance automatique qui considère « Saint-François » comme une personne, mais également un nettoyage des données qui n'a pas complètement corrigé les occurrences faisant référence aux Français. On voit ici les limites d'une telle technique de reconnaissance automatique du contenu d'un large corpus de données, ou plutôt de ces capacités de généralisation de telles méthodes sans prendre en compte les contextes spécifiques.

La troisième entité reconnue est celle du nom de famille « Rosset » qui peut faire référence à de nombreuses personnes différentes dans les cahiers Polier, ce nom de famille ayant fréquemment cour à Lausanne au XVIIIe siècle. En effet, sans prétendre à l'exhaustivité, on retrouve notamment la mention de Jean Alphonse Rosset, professeur, Pierre Antoine Louis

TABLEAU 2 – LES ENTITÉS DE PERSONNES ET DE LIEUX LES PLUS FRÉQUENTES

	Personnes		Lieux	
	Entité	**Fréquence**	**Entité**	**Fréquence**
1	mr le bailli	1546	lausanne	4288
2	francois	705	vernand	1618
3	rosset	315	paris	1279
4	vullyamoz	284	geneve	887
5	mr de vernand	284	la haye	546
6	polier	270	france	535
7	blanchard	266	bottens	508
8	baud	255	morges	447
9	tissot	246	hollande	412
10	blondel	240	londres	389

Rosset-Ginging, boursier, conseiller à Lausanne et banneret de la Palud ou encore Richard Rosset, justicier.

En ce qui concerne « Vullyamoz », il semble ici également que la reconnaissance automatique d'écriture manuscrite ait failli à distinguer « Vuilliamoz » de « Wullyamoz » et les autres orthographes possibles de ce patronyme. Cela peut notamment s'expliquer par la rareté de la lettre « W » dans notre ensemble de données d'entraînement. Parmi les différentes personnalités qui portent ce nom, il semble que l'on retrouve un pasteur, un qui fut un temps président du tribunal de la rue de Bourg ou encore Pierre Moïse Wullyamoz qui était conseiller à Lausanne.

La quatrième entité, « Mr de Vernand », fait généralement référence à son frère Georges Louis lorsque celui-ci ne s'adresse pas directement à celui-ci. On le voit notamment dans le cahier 31 lorsque le lieutenant baillival écrit « payer par cette premiere de change a l'ordre de mr de vernand colonel d'infanterie et capitaine commandant au regiment des gardes

suisses »[151]. De même la cinquième entité « Polier » nous témoigne de sa fréquence de correspondance avec diverses membres de sa famille du côté paternel.

La fréquence du nom « Blondel » est majoritairement due à Ferdinand Blondel qui était en charge des vignes du domaine familial. Finalement, les quatre entités restantes, soit « Blanchard », « Baud » et « Tissot » nous laisse la marque des noms de famille courants à Lausanne à l'époque de Jean Henri Polier de Vernand. Notons tout de même la présence d'un « Baud » procureur à Lausanne.

Vis-à-vis des entités de lieux les plus fréquentes, il n'a pas surprenant de voir Lausanne apparaître significativement plus souvent que toutes autres entités similaires. La deuxième entité « Vernand » n'est pas non plus surprenante, étant donné les liens entre cette branche de la famille Polier avec l'ancienne enclave lausannoise[152] ou Jean Henri possède des terres et est seigneur.

Il en va de même avec l'occurrence « Bottens », la septième occurrence la plus fréquente. En effet, toute une autre branche de la famille Polier est la seigneurie de Bottens. Polier parle notamment du « petit de Bottens » qui s'avère être Henri Etienne Georges Fitz Roger de Polier, le futur premier préfet du Léman et son filleul : « je dis au Petit de Bottens que les Plantamour parloint avec la plus grande indiscretion & que je le sollicitait de m'autoriser à dire que c'étoit, mais il s'éloigna bien vite en s'écriant hi hi hi, on dit tant de chose, d'ou je compris le fait n'etoit que trop vrai »[153].

La fréquence des occurrences de « La Haye » et « Hollande », respectivement cinquième et neuvième place, s'explique également par des liens familiaux puisque son frère Georges Louis de Polier s'y trouve au sein du régiment des Gardes-Suisses.

La troisième occurrence la plus fréquente est celle de « Paris » et on retrouve également à la sixième place « France ». Ces références sont fréquemment en référence à des extraits de journaux que Jean Henri Polier retranscrit fréquemment dans ces cahiers. D'autres sont également liés à des lettres de change. Cela vaut également pour l'entité « Londres » qui est la dixième la plus fréquente.

---

151 CH ACV P Rene Monod 31 p. 20

152 Hubler Lucienne, « Vernand », *hls-dhs-dss.ch*, https://hls-dhs-dss.ch/articles/049612/2013-07-08/, consulté le 10.03.2023.

153 CH ACV P Rene Monod 136, p. 13

Quant à l'occurrence « Morges », elle est fréquemment citée dans diverses affaires judiciaires quant il spécifié qu'une personne est originaire de Morges ou encore lorsque Polier parle de vente de vin ou d'achat de nouvelles cultures.

TABLEAU 3 – LES ENTITÉS D'ORGANISATIONS ET DIVERSES

	Organisations		Divers	
	Entité	**Fréquence**	**Entité**	**Fréquence**
1	consistoire	4023	bise	1034
2	supreme	733	temoin	905
3	crutz	265	batz	508
4	chateau	253	messel	440
5	aumone	135	pre	264
6	charite	132	moitie	220
7	surete	130	girouette	178
8	baillive	122	illegitime	173
9	accouchee	119	echu	152
10	condamnee	108	ecrire	135

Le tableau 3 nous montre les occurrences les plus fréquentes qui ont été labellisées sous le titre « organisation » ou « divers ». Il s'agit des entités nommées identifiées par le modèle de traitement du langage naturel `spaCy` comme appartenant à une organisation ou à une entreprise pour le premier, et à toutes les autres entités nommées qui ne sont pas classées dans les autres entités pour le deuxième.

On constate que les deux premières occurrences d'organisation les plus fréquentes sont « Consistoire » et « Supreme », ce qui démontre l'importance de cet aspect dans les charges de la position de lieutenant baillival occupée par Polier. Comme nous l'avons vu précédemment, Jean Henri Polier de Vernand se voyait généralement céder la place de président du consistoire de Lausanne normalement occupée par le bailli. A cela s'ajoute le neuvième terme le plus fréquent, « acouchee » qui démontre la centralité des questions de grossesses hors mariage dans ce tribunal des mœurs de l'Ancien régime. Le dixième terme,

« condamnee », genré au féminin, vient également renforcer cet aspect tandis que « surete » s'ajoute au vocable judiciaire. À cela s'ajoute les entités diverses que sont « temoin », « illegitime », respectivement deuxième et huitième.

On constate également la présence d'entité monétaire ayant cours à Lausanne au XVIIIe siècle comme « crutz » cité à 265 reprises parmi les organisations, « batz » à 508 reprises et « echus » a 152 reprises parmi les entités diverses. On peut également observer la présence parmi les entités diverses de « messel » et « pre » à la quatrième et cinquième place qui peut être comprise comme un indicatif d'informations agricoles.

Parmi les entités d'organisation ou retrouve également la présence de « aumône » qui porte la trace de la fréquentation de Polier de l'Église, sachant que celui-ci écrivait régulièrement dans son mémorial les serments qui l'intéressait particuliers et notait scrupuleusement les dons faits à la fin de ceux-ci. On retrouve également l'occurrence « charité » qui fait à la fois référence à l'acte de donner, mais peut aussi être en rapport avec l'École de la Charité qui existait alors à Lausanne. Finalement, les indications « bise » et « girouette » sont porteuses de la trace des scrupuleux relevés météorologiques effectués par Polier.

De manière générale, une analyse de fréquence des entités nommées est un moyen d'entrer dans un large corpus de données au moyen d'un processus automatisé de compréhension de texte. Cette approche nous fournit des informations utilisées sur les lieux les plus fréquemment cités dans la source. Cela peut également être utile pour identifier des thèmes ou des personnages récurrents.

4.3. L'ANALYSE DES THÈMES

Pour poursuivre l'analyse des sujets récurrents des journaux personnels de Jean Henri Polier de Vernand, une autre analyse automatisée pertinente est la *Latent Dirichlet Allocation* (LDA) effectuée grâce au script présenté en annexe V. Il s'agit d'une technique de modélisation thématique utilisée pour l'analyse de grands corpus de textes et qui s'avère particulièrement utile pour les tâches de classification et d'identification des sujets contenus dans ceux-ci[154].

[154] Sugimoto Cassidy R. *et alii*, « The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation », *Journal of the American Society for Information Science and Technology*, vol. 62, no 1, 2011, p. 187

Le script que nous avons créé met également en place une liste de mots à ne pas prendre en compte en combinant une liste de mots en français courants avec une liste de mots spécifique à notre corpus de source analysé ici. Ce sont donc des mots fréquents qui ne sont pas significatifs pour la compréhension du texte, telle que des prépositions, des pronoms, des conjonctions, etc. ou encore pour être plus spécifique à notre source les mois de l'année par exemple.

Qui plus est, le modèle adopte une approche de modélisation appelée « sac de mots » (*bag-of-words* ou BoW en anglais) qui analyse les mots indépendamment les uns des autres, ce qui signifie que l'ordre des mots et des documents n'a pas de signification inhérente[155]. Dans la méthode LDA on extrait des thèmes où chacun est caractérisé par une distribution sur les mots avec des thèmes latents[156]. Cette approche permet de considérer que les mots peuvent se répéter dans plusieurs sujets et qu'un document peut contenir plusieurs sujets plutôt que de classer singulièrement la nature d'un document[157].

Le script permet dans un premier temps d'extraire le texte des transcriptions des cahiers Polier et utilise `spaCy` pour générés des *lemmas,* soit la racine des mots qui permet de regrouper des mots apparentés aux terminaisons différentes. Cela peut permettre de rassembler plusieurs données comme « réclamer » et « réclamation » ensemble et ainsi améliorer la performance du modèle et donc mieux reconnaître les sujets ou les thèmes sous-jacents au corpus de transcription des cahiers Polier.

Après avoir exclu les mots d'arrêts spécifiés et filtré les mots non alphabétiques, les résultats sont stockés dans une liste appelée « tokens ». Cette étape de tokenisation, soit le fait de diviser le texte en mots individuels selon la méthode des sacs de mots présentée ci-dessus. Les données textuelles extraites sont ensuite vectorisées en utilisant la fonction `CountVectorizer()` de la bibliothèque Python `scikit-learn`. Cela permet de convertir le texte extrait en données binaires avant de les soumettre à l'analyse de LDA.

L'algorithme procède ensuite en sélectionnant le vocabulaire à utiliser pour la LDA. Pour pouvoir fonctionner, celui-ci a besoin de savoir quels sont les termes qui sont utilisés dans les documents et leur fréquence d'apparition. Il commence donc par faire une liste de tous les mots dans les documents – soit la tokenisation – puis il ajoute un filtre supplémentaire. En

---

155 *Ibid.*
156 *Ibid.*
157 *Ibid.*

effet, tel que définit dans notre script en utilisant la notation [:15], la LDA ne va prendre en compte les 15 termes les plus fréquents du vocabulaire total afin d'éviter que les mots les plus fréquents ne soient surreprésentés dans les résultats de l'analyse et éviter de biaiser les résultats. De plus, nous excluons également les mots qui n'apparaissent pas au moins trois fois dans le corpus afin d'exclure les mots qui seraient trop rares pour être représentatifs. En sélectionnant ainsi le vocabulaire à prendre en compte dans l'analyse, on obtient ainsi les thèmes sous-jacents au corpus et un résultat plus représentatif.

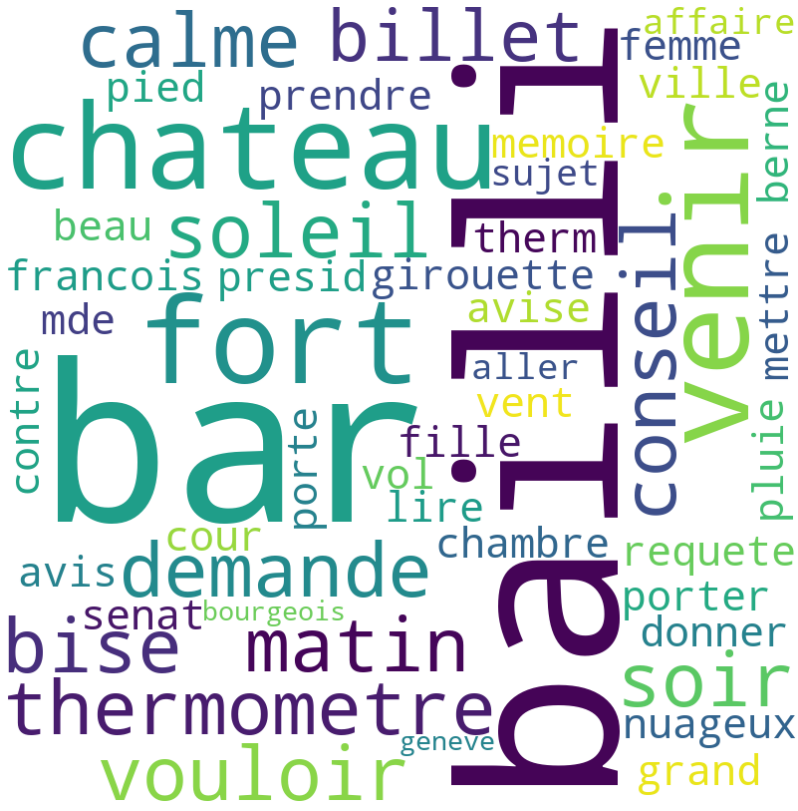
Il s'agit donc d'un algorithme de modélisation des sujets qui identifie les thèmes dans un ensemble de documents en attribuant un poids à chaque mot défini par sa fréquence d'apparition vis-à-vis de chaque cluster. L'algorithme nous fournit donc une liste de mot par « thème » identifié auquel nous devons ensuite trouver un titre ou du moins une appellation permettant de comprendre leur contenu. On obtient ainsi dix groupes de mots les plus représentatifs par le modèle LDA. Cependant, nous avons réuni certains groupes ensemble et sommes arrivés à un total de sept thèmes, estimant que certains touchaient à des sujets similaires.

Afin de visualiser les données générées, nous avons choisi de nous tourner vers la représentation sous forme de nuage de mots, qui permet à la fois de visualiser les mots les plus représentatifs de chaque cluster, mais également de mettre en lumière leur poids relatif au sein de ce cluster. Il permet la visualisation des mots les plus représentatifs de manière claire et facile à interpréter, notamment pour identifier les thèmes ou sujets communs.

Pour chaque sujet, nous avons également chercher à savoir quels étaient les mots les plus caractéristiques de chaque sujet et bien que ceux-ci ne semblent pas toujours faire sens ils peuvent nous aider à comprendre à quel thème général ce rattache chaque groupe détecté par la LDA. Ceux-ci sont des mots spécifiques à chaque sujet, c'est-à-dire qu'ils sont rarement fréquents dans un autre groupe détecté par la LDA et permettent donc de mieux catégoriser un sujet. Cependant, ceux pouvant être rare dans le corpus, ils peuvent avoir échappé à l'étape de nettoyage des données et donc ne sont pas systématiquement présentés ici.

4.3.1. Les notes quotidiennes

TOPIC 1



TOPIC 2

Le premier thème que nous avons identifié réunit les deux premiers groupes détectés par la LDA puisque ceux-ci semblent tous deux toucher aux notes quotidiennes du lieutenant baillival. En effet, on y retrouve des thèmes variés tels que les relevés météorologiques quotidiens effectués par le Lausannois, ainsi que des informations touchant à la gestion des transactions financières. Les mots les plus représentatifs du premier cluster incluent «sol», «soleil», «livre», «bar», «beau», «quittance», «girouette», «recu», «payer», «thermometre», «batz», «demi», «therm», «bise», «vent», «pluie», «paye», «grand», et «barometre».

Dans un même temps, le deuxième cluster on retrouve également «bar», «thermometre», «fort», «calme», «soleil», «bise», «matin», «soir», «girouette» ou «vent». Ceux si se rapprochent donc du premier, même si d'autres termes se rapprochent plus des occupations liées à ces mandats. En effet on retrouve

notamment «bailli», «château», «demande», «vouloir», «conseil» , «billet» et «porter».

Ces mots semblent liés à ces événements quotidiens tels que les visites au château où réside le bailli ou encore où se tiennent certaines cour judiciaire mais aussi des termes faisant penser à des discussions et aux conseils de la ville de Lausanne.

Pour toutes ces raisons, nous avons regroupé ces deux résultats de la LDA ensemble sous la qualification des événements du quotidien. Ceux-ci ont peu à nous apprendre sur le contexte lausannois, mais nous démontre avec quelle acuité Jean Henri Polier de Vernand transcrivait son quotidien.

4.3.2. Les charges de lieutenant baillival

TOPIC 3

Le deuxième thème que nous avons identifié touche plus clairement aux tâches qui incombent à un lieutenant baillival. En effet, les mots de ce cluster se concentrent sur des termes liés à la ville de Lausanne et son administration, ainsi que sur les relations avec la seigneurie et Berne. Les mots les plus représentatifs incluent « ville », « seigneur », « date », « bailli », « honneur », « berne », « chambre », « conseil », « enfant », « donner », « seigneurie », « avoir », « consistoire », « signe », « bon », « ordre », « contre », et « sujet ».

Les termes les plus fréquents, tels que "ville", "seigneur", "donne" et "conseil", reflètent la nature administrative et politique du topic. Les sujets semblent ici encore relativement généraux, d'autant que certains de ces aspects comme le consistoire reviennent plus spécifiquement dans un autre cluster. Ici on se trouve plus proche des aspects liés aux

pouvoirs politique. Dans les mots les plus caractéristiques, on retrouve notamment « Illustre » et « République », qui indique que ce thème est fortement lié aux relations avec Berne.

4.3.3. Les affaires du Consistoire

TOPIC 4

Ce groupe de mots détectés par la LDA est fortement lié au consistoire et aux différentes affaires qui y sont liées. En effet, on retrouve des termes tels que « enfant », « fils » ou « fille » qui peuvent potentiellement être en relation avec des affaires de paternité. On retrouve également le vocabulaire du statut marital avec « mariage », « marier », mais aussi « veuve ». L'idée de « demande », « mandat », « rapport » et « copie » quant à eux se rapprochent plus de la thématique de la loi. Une autre occurrence est celle de « tuteur » qui évoque des relations de tutelle et de protection, impliquant souvent des décisions et des mandats de la part des autorités religieuses.

Qui plus est on retrouve parmi les mots les plus caractéristiques du topic les termes « charnel » à 239 reprises, « interpelle » à 92 reprises et « encint » à 44 reprises. Si ce dernier est manifestement dû à une mauvaise reconnaissance de l'écriture manuscrite qui a échappé à l'étape du nettoyage des données, sinon quoi celui-ci aurait été exclu du vocabulaire utilisé pour la LDA afin d'obtenir les thèmes sous-jacents, sa présence nous confirme l'idée qu'un des sujets centraux vis-à-vis des consistoires touche aux questions de paternité et d'enfants hors mariage. Parmi les autres mots les plus caractéristiques on retrouve également « juree » à 92 reprises, « recidive » à 28 reprises ainsi que « ajourne » à 23 reprises, nous ramenant ici encore à des aspects judiciaires.

4.3.4. La religion

TOPIC 5

Ce thème se rapproche plus de la question de la religion et est principalement axé sur des notions telles que l'homme, la loi, le roi, la vertu, l'esprit et le pouvoir. Jean Henri Polier de Vernand était très attaché aux questions religieuses et on le voit notamment avec des mots tels que « morale », « genie », « richesse », « sublime », « bienfaisance », « monarque » et « nature » qui suggère des considérations morales et politiques en lien avec la religion avec les occurrences « dieu », « vertu ».

Le mot « homme » est le plus fréquent dans ce topic, ce qui renforce l'idée d'une réflexion sur l'être humain. La « loi » et le « roi » peuvent évoquer le pouvoir et l'autorité, tandis que « vertu » et « esprit » peuvent être compris comme des qualités morales et intellectuelles.

Les mots les plus caractéristiques du topic #8 comprennent « morale », « genie », « empire », « richesse », « bienfaisance » et « monarque ». Ces termes semblent se concentrer sur le rôle de l'État et de la société dans la promotion de la vertu et de la morale. Le mot « monarque » peut être interprété comme une réflexion sur les systèmes politiques et la relation entre le pouvoir et la vertu.

4.3.5.Correspondance formelle

TOPIC 6

TOPIC 7

Les clusters 6 et 7 ont été réunis dans un même thème appelé correspondance formelle, qui tient beaucoup au style d'écriture de Jean Henri Polier de Vernand. En effet, si celui-ci utilise un ton usuel pour sa correspondance avec ses proches comme son frère par exemple, que l'on retrouve par ailleurs avec le terme « Haye » dans le topic 7. « sa langue est tout autre pour ses lettres de félicitations ou de condoléances qui ressemblent à des "exercices de style" aux phrases apprêtées »[158].

Ainsi on retrouve des termes similaires à deux utilisés dans le contexte de la religion, avec notamment « roi », « Berne », « homme » ou « pouvoir ». Des termes plus génériques, mais qui correspondent à un style littéraire plus sophistiqué qu'il utilise dans sa correspondance officielle comme « cher », « vouloir », « bon » et « grand ». Les termes comme « apprendre », « remercier », « souhaite » et « prier » évoque également l'idée

[158] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op. cit*, p. 56

d'une correspondance. Les mots les plus caractéristiques reflètent des thèmes plus spécifiques liés aux personnes impliquées dans la correspondance, tels que des noms de personnes ou des termes amicaux comme « affectueusement » ou de doléance comme « regret ».

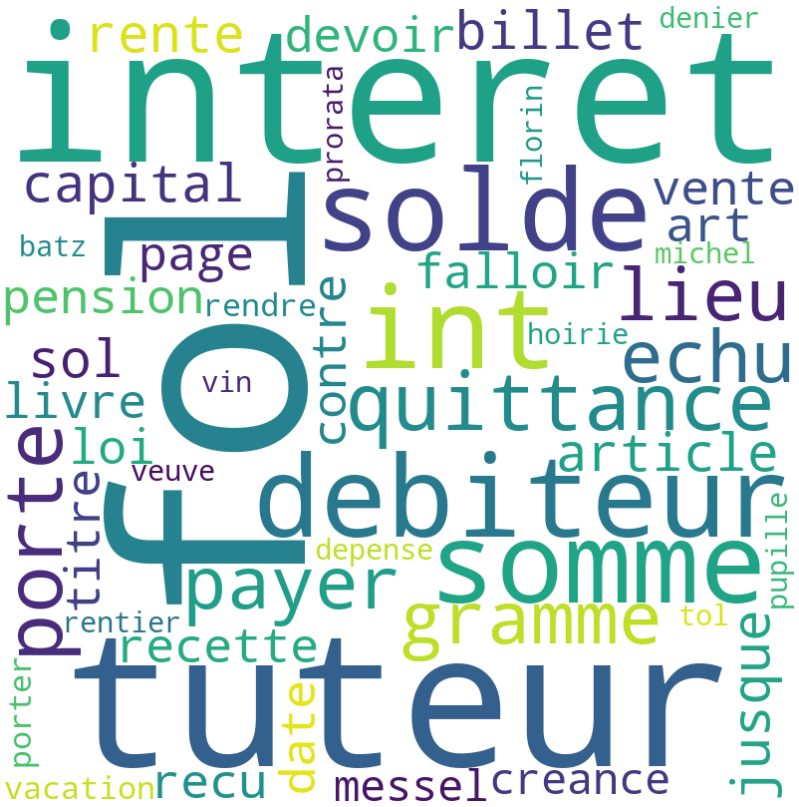
4.3.6. Le domaine de Vernand

TOPIC 8

Le sixième thème que nous avons identifié est lié aux terres agricoles de la famille Polier de Vernand et notamment ses vignes. On voit notamment apparaître le nom « Blondel » qui était en charge des vignes du lieutenant baillival au Mont, mais également « Romand » qui fait référence à Mathieu Romand qui s'occupait de celles situées à Allaman. Les deux localités se retrouvent parmi les mots les plus représentatifs du cluster.

Le thème en général se centre autour de la production de vin et tout ce qui touche à la viticulture. On retrouve notamment les mots tels que « pot », « vin », « vigneron », « vigne », « vendange » et « prix ». La production de vin était l'une des sources principales de revenus pour Jean Henri Polier. Le mot "partisseur" indique également que ce topic pourrait être lié à la distribution ou à la vente de vin. On peut également remarquer la présence du mot "fuste", qui fait référence aux barils utilisés pour le vieillissement du vin.

4.3.7. Le mémorial comme livre de comptes



TOPIC 9

TOPIC 10

Le dernier sujet qui nous intéresse ici touche principalement aux transactions financières, que ce soit les achats quotidiens ou la gestion financière des biens de Polier de Vernand. Ce dernier avait pour habitude de noter toutes ses dépenses. En effet, on retrouve dans les cahiers Polier beaucoup de référence à des prêts et à des lettres de change et «comme à cette époque les banques n'étaient pas encore les établissements de crédit qu'elles sont devenues et que les actions de sociétés n'existaient pour ainsi dire pas, si l'on excepte celles de la Compagnie des Indes orientales, par exemple, les prêts se faisaient de particulier à particulier, entre gens qui se connaissaient ou qui pouvaient fournir des garanties suffisantes»[159].

Les mots les plus représentatifs du topic 9 sont notamment « debiteur », « somme », « payer » et « capital » qui suggèrent des transactions

[159] Morren Pierre, *La vie lausannoise au XVIIIe siècle…, op. cit*, p. 72

financières et des prêts. On constate également la présence de « tuteur » dans ce les mots les plus représentatifs qui était également présents dans le topic touchant au consistoire, ce qui n'est pas étonnant vu la nature de la fonction.

Le topic 10 quant à lui touche plus aux transactions financières du quotidien. Il est caractérisé par des mots tels que « sac », « caisse », « debit », « payer », « venir », qui peuvent suggérer des transactions commerciales ou des échanges de biens.

4.4. DIFFICULTÉS RENCONTRÉES

De notre point de vue, un travail d'histoire touchant au domaine des humanités numériques permet à la fois de constater l'impact significatif de l'apprentissage automatique sur notre capacité d'élaboration et de compréhension de grands corpus de données historique, mais également les difficultés techniques liées à l'utilisation de ces méthodes qui peuvent constituer des obstacles importants à leur mise en œuvre.

Parmi les défis qui ont traversé ce travail, on retrouve notamment des incompatibilités entre la nécessité d'utiliser Tensorflow pour exécuter les nombreux scripts évoquée dans ce travail et qui s'avère incompatible avec l'ordinateur principalement utilisé dans ce travail, soit un Mac doté d'une puce M1 d'Apple.

TensorFlow est une bibliothèque logicielle open source largement utilisé pour l'apprentissage automatique, développé par Google, qui permet aux chercheurs de construire, d'entraîner et de déployer des modèles d'apprentissage automatique[160]. Malheureusement, l'architecture de la puce M1 n'est actuellement pas prise en charge par TensorFlow, ce qui signifie que les chercheurs qui ont investi dans un Mac équipé d'une puce M1 risquent de ne pas pouvoir utiliser TensorFlow pour leurs recherches.

Si le monde de l'intelligence artificielle et de l'apprentissage automatique progresse très rapidement, c'est également le cas pour la demande de matériel et de logiciels puissants pour les prendre en charge, mais ceux-ci ne sont pas nécessairement développés au même rythme. Ainsi l'un des aspects des humanités numériques est également la dépendance à des technologies qui sont hors de son contrôle.

Pour résoudre ce problème, il nous a fallu créer sur notre ordinateur un environnement virtuel afin d'y faire fonctionner TensorFlow. La création d'un environnement virtuel pour

[160] IA School, « Tensorflow : tout connaitre sur cet outil Open Source IA School », *IA School*, https://www.intelligence-artificielle-school.com/les-technologies/tensorflow/, consulté le 10.03.2023.

exécuter TensorFlow sur une puce M1 nécessite un temps considérable et des connaissances techniques avancées qu'il n'est pas aisé de maîtriser.

Passé cet obstacle, un autre défi important a été le temps de traitement requis pour l'élaboration du modèle de reconnaissance d'écriture manuscrite. Bien que ces modèles soient une ressource importante pour les historiens qui travaillent avec des documents historiques manuscrits, permettant de transcrire et d'analyser le texte contenu dans ces mêmes documents, le temps requis pour leur élaboration s'avère être important. La retranscription des données qui ont servi à l'entraînement du modèle, leur exportation et leur transformation dans un format utilisable pour HTR-Flor++ représentent une temporalité importante.

Qui plus est, l'utilisation d'HTR-Flor++ en soit représente un modèle d'apprentissage automatique sophistiqué qui nécessite des calculs importants et exigeants en termes de puissance de traitement et de mémoire, ce qui rend leur utilisation sur des ordinateurs locaux. Pour résoudre ce problème, nous avons dû nous tourner vers la plateforme Google Colab qui à son tour présente également ses propres défis.

En effet pour résoudre le problème des ordinateurs personnels qui ne peuvent avoir suffisamment de ressources pour exécuter efficacement ces modèles d'apprentissage en profondeur, de nombreux chercheurs se tournent vers des plateformes cloud telles que Google Colab pour pouvoir exécuter ces modèles complexes. Google Colab permet d'accéder à un environnement de développement en ligne doté de ressources matérielles importantes, y compris des processeurs graphiques haut de gamme (GPU) et des unités de traitement de tenseur (TPU)[161].

Cependant cela implique également d'être dépendant de cette plateforme qui met en place des limitations en termes de stockage de données ou de temps d'exécution. Cela a ralentit le temps de création des prédictions des cahiers Polier, que ce soit par la nécessité d'attendre que les fichiers soient disponibles sur le Google Drive — qui s'avérait également nécessaire — au stockage limité, mais également de prendre son mal en patience lorsque la plateforme de l'entreprise Alphabet imposait une limite d'utilisation. Qui plus est, cela implique également des défis en termes de sécurité des données et de dépendance à une plateforme tiers.

---

[161] Sousa Neto Arthur Flor de et alii, « HTR-Flor++: A Handwritten Text Recognition System… », *art. cit.,* p. 3

Nous avons également vu que l'étape du nettoyage des données est non seulement cruciale, mais également ardue sur un si grand corpus de données qui regroupe à la fois des données météorologiques, des listes d'achats, des décomptes d'intérêts, des notes sur les sermons tenus à l'église, des comptes rendus de séances liés à un mandat baillival ou encore des brouillons de correspondances à la fois privée et professionnelle et donc une grande variété de formats et de structures.

Une fois ces données nettoyées et préparées, l'analyse de leur contenu comme nous l'avons fait ici avec une LDA nécessite des calculs complexes coûteux en temps et en ressources pour pouvoir extraire des thèmes et des sujets significatifs sur un grand corpus. Qui plus est, ces scripts ne sont pas aisés à rédiger et à adapter à nos données.

Bien que ces difficultés techniques puissent être frustrantes, les avantages de l'utilisation de l'apprentissage automatique dans la recherche historique restent présents. En effet, l'apprentissage automatique peut nous permettre d'analyse et de comprendre des données historiques d'une manière qui était auparavant impossible. En tant que chercheur en histoire et en humanités numériques, il nous incombe de trouver des moyens de surmonter ces difficultés techniques et d'exploiter la puissance de l'apprentissage automatique au profit de la recherche historique afin d'exploiter son potentiel et de continuer à apporter des contributions significatives à notre compréhension du passé.

CONCLUSION

Dans ce travail, nous avons établi la position sociale, politique et administrative que Jean Henri Polier de Vernand a occupée à Lausanne au fil de sa vie. Ces cahiers qui constituent son mémorial allant de mai 1754 à septembre 1759 – à l'exception de quelques omissions qui se sont perdus – nous ont donné à voir sa méticulosité à consigner ses comptes, ses notes viticoles et agricoles, ses relevés météorologiques et ses observations sur son quotidien en tant que lieutenant baillival. On y trouve les résumés de ses journées, y compris des séances et des réunions auxquelles il participait avec ses nombreux mandats et une importante correspondance avec divers membres de sa famille et officiels en place au cours de sa rédaction.

Désormais, les 26'300 pages qui constituent cette source ont été retranscrites au format numérique. Avec ce travail, nous espérons que cette source – bien que déjà connue, mais rarement utilisée dans son intégralité – pourra susciter de nouvelles recherches, notamment vis-à-vis des relevés météorologiques précis ou des observations sur la vie sociétale lausannoise. Les fichiers issus de ce travail ont été remis au projet *Lausanne Time Machine*.

Si nous espérions aux débuts de ce travail effectué nous-mêmes un travail historique plus poussé sur cette source si riche, les difficultés techniques pour créer un modèle de reconnaissance automatique d'écriture manuscrite spécifique à un document historique datant du XVIII^{e} siècle nous ont poussés à voir nos ambitions à la baisse. Cependant, ces embûches ont également permis de mettre en avant les difficultés que peuvent représenter les humanités numériques pour quelqu'un qui n'a pas une formation complète en informatique.

La plateforme la plus adaptée à un public non initié n'étant pas accessible financièrement pour l'envergure de ce travail, il nous a fallu apprendre à maîtriser HTR-Flor++, un programme entièrement rédigé en langage Python. Avec cette méthode, nous sommes parvenus à un taux d'erreurs acceptable pour extraire de la source physique une transcription numérique des cahiers de Jean Henri Polier de Vernand.

Grâce à des techniques de reconnaissance automatique du contenu d'un document, nous avons pu nous faire une idée des personnalités, lieux et entités diverses les plus importantes du corpus. Nous avons également pu extraire du texte les aspects principaux du texte rédigé

sur une quarantaine d'années par le lieutenant baillival. Confrontés à la littérature existante et aux parties plus biographiques sur la vie de Polier de Vernand, nous avons ainsi trouvé un point d'entrée viable dans un tel corpus de données et trouvé des résultats significatifs au regard du contexte exposé.

Le développement des humanités numériques nous permet désormais de nouvelles possibilités de recherches, même si celles-ci ne restent pas sans poser leurs défis spécifiques. Parmi ces possibilités, la reconnaissance automatique d'écriture manuscrite permet de rendre accessible des sources jusque là laborieuses à étudier. Celle-ci est néanmoins une technologie qui peut s'avérer complexe et coûteuse à mettre en place. Qui plus est, les historiens doivent se familiariser avec des techniques computationnelles complexes et peuvent passer des heures à configurer leur modèle et les adapter à leur corpus.

De plus, les projets de reconnaissance automatique d'écriture manuscrite peuvent être confrontés à des problèmes de qualité des données. Les documents historiques peuvent être en mauvais état, comporter des tâches ou encore des difficultés inhérentes à la nature humaine de l'écriture manuscrite qui peuvent compliquer sa reconnaissance par un ordinateur.

Dans ce contexte, les aspects propres au traitement des données peut prendre le pas sur les aspects proprement historiques. Cependant, nous postulons que l'accessibilité à des sources aussi importante reste un résultat historique en soi. Ces technologies relativement nouvelles offrent de nombreuses possibilités pour la recherche historique. Pour autant, il importe également de souligne que celles-ci ne sont pas parfaites et implique de passer du temps à vérifier et à corriger les erreurs du modèle HTR, un tâche qui peut s'avérer chronophage et, surtout lorsque réalisée individuellement, pesé sur les objectifs de la recherche.

SOURCES

ACV, «Archives cantonales vaudoises: rapport d'activités 2021 by État de VAUD — Issuu», 12.08.2022, https://issuu.com/etatdevaud/docs/rapport-annuel-2021, consulté le 06.03.2023.

ACV P René Monod 1-219 Livre de raison, soit «Mémorial universel...» de Jean-Henri Polier de Vernand, lieutenant baillival du 1er mai 1754 au 2 février 1791

FIGURES

PROGRAMME

FLOR ARTHUR, arthurflor23/handwritten-text-recognition, 08.03.2023 [14 avril 2019], Python, MIT License, https://github.com/arthurflor23/handwritten-text-recognition.

BIBLIOGRAPHIE

SUR JEAN-HENRI POLIER DE VERNAND

ABETEL Emmanuel, «Polier de Vernand, Jean-Henri», *hls-dhs-dss.ch*, https://hls-dhs-dss.ch/articles/017839/2009-04-20/, consulté le 03.03.2023.

CHARRIÈRE DE SÉVERY W., «Le cercle de la rue de Bourg fondé en 1761», *Société vaudoise d'histoire et d'archéologie*, vol. 22, n°8, 1914, p. 250-254, DOI: 10.5169/SEALS-19506.

DELÉDEVANT Henri *et alii*, *Le livre d'or des familles vaudoises : répertoire général des familles possédant un droit de bourgeoisie dans le canton de Vaud,* Lausanne : Ed. Spes, 1923, 435 p.

FAVEZ Valérie, *Etude du «Mémorial universel» tenu par Jean-Henri Polier de Vernand, lieutenant baillival: gestion d'un patrimoine (1754 à 1761),* Mémoire de Master*,* Université de Lausanne, 1991, 122 p.

HUBLER Lucienne, «Vernand», *hls-dhs-dss.ch*, https://hls-dhs-dss.ch/articles/049612/2013-07-08/, consulté le 10.03.2023.

KIENER Marc, *Dictionnaire des professeurs de l'Académie de Lausanne (1537-1890)*, Lausanne : Université de Lausanne, 2005, 689 p.

MONTET Albert, *Dictionnaire biographique des Genevois et des Vaudois: qui se sont distingués dans leur pays ou à l'étranger par leurs talents, leurs actions, leurs œuvres littéraires ou artistiques, etc*, Lausanne : G. Bridel, 1877-1878, 445 p.

MORREN Pierre, *La vie lausannoise au XVIIIe siècle: d'après Jean-Henri Polier de Vernand, lieutenant baillival,* Genève : Labor et Fides, 1970, 622 p.

SUR LES ÉGOCUMENTS ET BIOGRAPHIES

GINZBURG Carlo, *Le fromage et les vers: l'univers d'un meunier du XVIe siècle*, Paris : Flammarion, 1980, 220 p.

TOSATO-RIGO Danièle, «Pratiques de l'écrit et histoire par la marge: autour des "egodocuments" en Suisse romande au XVIIIe siècle», *Verlag Karl Schwegler AG*, vol. 67, n° 4, 2010, p. 261-268, DOI: 10.5169/SEALS-169847.

TOSATO-RIGO Daniele, «L'archive privée: au coeur des pratiques sociales et culturelles», *arbido*, https://arbido.ch/fr/edition-article/2013/privatarchive/larchive-privée-au-coeur-des-pratiques-sociales-et-culturelles, consulté le 15.02.2023.

SUR L'HISTOIRE DU CANTON DE VAUD ET SON SYSTÈME JUDICIAIRE

ANTOINE Samuel, «Consistoire, Conseil des XXIV et police des mœurs au XVIIIe siècle: les autorités lausannoises face aux "filles de mauvaise vie"», *Revue historique vaudoise*, vol. 118, Société vaudoise d'histoire et d'archéologie, vol. 118, 2010, p. 123-134, DOI: 10.5169/seals-847042.

BIAUDET Jean-Charles *et alii*, *Histoire de Lausanne*, Toulouse : Privat, 1982, 456 p.

CHUARD Corinne, *Histoire vaudoise: un survol*, Gollion : Infolio, 2019, 159 p.

JACKSON Jeremy Charles, *The Evolution of a Municipal Oligarchy: Lausanne, 1536-1798*, Ann Arbor Mich : Univ. Microfilms international, 1977, 255 p.

MAILLEFER Paul, *Histoire du Canton de Vaud dès les origines*, Lausanne : Payot, 1903, 553 p.

MATZINGER-PFISTER Regula, «L'introduction des consistoires dans le Pays de Vaud», in TOSATO-RIGO Daniele et STAREMBERG GOY Nicole (dir.), *Sous l'oeil du consistoire. Sources consistoriales et histoire du contrôle social sous l'Ancien Régime*, Etudes de Lettres, 2004, p. 113-124.

SALVI Elisabeth, «La justice de LL. EE. au siècle des Lumières», in FLOUCK François *et alii* (dir.), De l'ours à la cocarde: régime bernois et révolution en pays de Vaud (1536-1798), Lausanne : Editions Payot, 1998, p. 325-345.

SPALINGER René, *Quand Mozart passait à Lausanne: chronique inédite*, Genève : Slatkine, 2006, 447 p.

STAREMBERG GOY Nicole, « Contenir la parole et le geste à Lausanne au XVIIIe siècle. Le Consistoire de la Ville face à la violence», in TOSATO-RIGO Daniele et STAREMBERG GOY Nicole (dir.), *Sous*

l'oeil du consistoire. Sources consistoriales et histoire du contrôle social sous l'Ancien Régime, Etudes de Lettres, 2004, p. 175-192.

TOSATO-RIGO Daniele et STAREMBERG GOY Nicole, «Avant-propos», in TOSATO-RIGO Daniele et STAREMBERG GOY Nicole (dir.), *Sous l'oeil du consistoire. Sources consistoriales et histoire du contrôle social sous l'Ancien Régime*, Etudes de Lettres, 2004, p. 5-12.

SUR LES HUMANITÉS NUMÉRIQUES ET L'HISTOIRE

BEAUDE Boris, «(re)Médiations numériques et perturbations des sciences sociales contemporaines», *Sociologie et sociétés*, vol. 49, n° 2, Les Presses de l'Université de Montréal, 2017, p. 83-111, DOI: 10.7202/1054275ar.

BURNARD Lou, «La TEI et le XML», in *Qu'est-ce que la Text Encoding Initiative?*, Marseille : OpenEdition Press, 2015, http://books.openedition.org/oep/1298.

CARDON Dominique, *Culture numérique*, Paris : Presses de Sciences Po, 2019, 432 p.

EDELSTEIN Dan, « Intellectual History and Digital Humanities », *Modern Intellectual History*, vol. 13, n° 1, Cambridge University Press, 2016, p. 237-246, DOI: 10.1017/S1479244314000833.

EVANS James A. et ACEVES Pedro, « Machine Translation: Mining Text for Social Theory », *Annual Review of Sociology*, vol. 42, n° 1, 2016, p. 21-50, DOI: 10.1146/annurev-soc-081715-074206.

FOLSOM Ed, « Database as Genre: The Epic Transformation of Archives », *Pmla-publications of The Modern Language Association of America*, vol. 122, 2007, p. 1571-1579, DOI: 10.1632/pmla.2007.122.5.1571.

GULDI Jo et ARMITAGE David, *The History Manifesto*, Cambridge : Cambridge University Press, 2014, 165 p.

HOCKEY Susan, « The History of Humanities Computing », in *A Companion to Digital Humanities*, John Wiley & Sons, Ltd, 2004. *En ligne :* https://companions.digitalhumanities.org/DH/?chapter=content/9781405103213_chapter_1.html

KLEIN Lauren et EISENSTEIN Jacob, « Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives », *Scholarly and Research Communication*, vol. 4, n° 3, 2013, DOI: 10.22230/src.2013v4n3a121.

SUR LA RECONNAISSANCE AUTOMATIQUE DE TEXTE MANUSCRIT

DE SOUSA NETO Arthur Flor *et alii*, «HDSR-Flor: A Robust End-to-End System to Solve the Handwritten Digit String Recognition Problem in Real Complex Scenarios», *IEEE Access*, vol. 8, 2020, p. 208543-208553, DOI: 10.1109/ACCESS.2020.3039003.

DE SOUSA NETO Arthur Flor de *et alii*, « HTR-Flor: A Deep Learning System for Offline Handwritten Text Recognition », in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, p. 54-61, DOI: 10.1109/SIBGRAPI51738.2020.00016.

DE SOUSA NETO Arthur Flor de *et alii*, « HTR-Flor++: A Handwritten Text Recognition System Based on a Pipeline of Optical and Language Models », in *Proceedings of the ACM Symposium on Document Engineering 2020*, New York, NY, USA : Association for Computing Machinery, 2020, p. 1-4, DOI: 10.1145/3395027.3419603.

DE SOUSA NETO Arthur Flor de *et alii*, « Towards the Natural Language Processing as Spelling Correction for Offline Handwritten Text Recognition Systems », *Applied Sciences*, vol. 10, no 21, Multidisciplinary Digital Publishing Institute, 2020, p. 7711, DOI: 10.3390/app10217711.

DE SOUSA NETO Arthur Flor de *et alii*, «A robust handwritten recognition system for learning on different data restriction scenarios», *Pattern Recognition Letters*, vol. 159, 2022, p. 232-238, DOI: 10.1016/j.patrec.2022.04.009.

GATOS Basilis *et alii*, « Segmentation of Historical Handwritten Documents into Text Zones and Text Lines », in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Greece : IEEE, 2014, p. 464-469, DOI: 10.1109/ICFHR.2014.84.

MUEHLBERGER Guenter *et alii*, « Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study », *Journal of Documentation*, vol. 75, n° 5, Emerald Publishing Limited, 2019, p. 954-976, DOI: 10.1108/JD-07-2018-0114.

NOCKELS Joe *et alii*, « Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research », *Archival Science*, vol. 22, n° 3, 2022, p. 367-392, DOI: 10.1007/s10502-022-09397-0.

PURCELL Jake, « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, vol. 7, n° 13, 2021, p. 13, DOI: 10.5334/johd.46.

SANCHEZ Joan Andreu *et alii*, « ICFHR2014 Competition on Handwritten Text Recognition on Transcriptorium Datasets (HTRtS) », in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Greece : IEEE, 2014, p. 785-790, DOI: 10.1109/ICFHR.2014.137.

SCHEIDL Harald, *Handwritten Text Recognition in Historical Documents*, Mémoire de master, Vienne : Technische Universität Wien), 2018, 80 p.

SUR L'ANALYSE AUTOMATISÉ DU CORPUS

NOUVEL Damien et alii, *Les entités nommées pour le traitement automatique des langues*, ISTE Group, 2015, 169 p.

OTHMEN Dhifallah, «Extraction et entraînement des entités nommées avec spaCy», *Medium*, 18.12.2019, https://medium.com/extraction-et-entrainement-des-entites-nommees-avec-spacy, consulté le 08.04.2023.

IA School, « Tensorflow: tout connaitre sur cet outil Open Source IA School », *IA School*, https://www.intelligence-artificielle-school.com/les-technologies/tensorflow/, consulté le 10.03.2023.

SUGIMOTO Cassidy R. *et alii*, « The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation », *Journal of the American Society for Information Science and Technology*, vol. 62, n° 1, 2011, p. 185-204, DOI: 10.1002/asi.21435.

ANNEXES

ANNEXE I : CONVERSION DES EXPORTATIONS DE TRANSKRIBUS POUR HTR-FLOR++

```
from lxml import etree as letree
import cv2
import tqdm
import glob
import os
import random
import numpy as np
```

```
from PIL import Image, ImageDraw
from scipy.ndimage import rotate

def baseline2box(pts, up: int = 20, down: int = 5, margin_l: int = 5,
margin_r: int = 5):
    min_y = np.min(pts[:, 0])-margin_l
    max_y = np.max(pts[:, 0])+margin_r
    min_x = np.min(pts[:, 1])-up
    max_x = np.max(pts[:, 1])+down

    return min_y, max_y, min_x, max_x


def checkCoords(min_x, max_x, min_y, max_y):
    return int(np.max([min_x, 0])), max_x, int(np.max([min_y, 0])), max_y


def filterImage(gray_image):
    fimage = cv2.bilateralFilter(gray_image, 7, 50, 50)
    value = np.ravel(fimage)
    paper, ink = np.percentile(value, 95), np.percentile(value, 5)

    white_fimage = (fimage-ink)/(paper-ink)*255
    white_fimage[white_fimage > 255] = 255
    white_fimage[white_fimage < 1] = 0
    white_fimage = white_fimage.astype('uint8')

    return white_fimage
```

```
def formatList(list_):
    string = ''
    for item in list_:
        string += f'{item}\n'
    return string
```

```
pad = 60

list_segments = []
list_notebooks = []

for path in tqdm.tqdm(sorted(glob.glob('...DATA/*.jpg'))):



        # Load image and recover its name
        image = filterImage(cv2.imread(path, 0))
        image_name = path.split("/")[-1][:-4]
        notebook, page_n = image_name.split('_')[-1].split('-')

        # Create directory for each notebook
        os.makedirs(f'{notebook}/BenthamDatasetR0-GT/Images/Lines/',
         exist_ok=True)
        os.makedirs(f'{notebook}/BenthamDatasetR0-GT/Partitions/',
         exist_ok=True)
        os.makedirs(f'{notebook}/BenthamDatasetR0-GT/Transcriptions/',
         exist_ok=True)



        # Read Transkribus XML output
        tree = letree.parse(
            f'{path.split("/CH_ACV")[0]}/CH_ACV_P_RENE_MONOD_{notebook}/
         page/CH_ACV_P_RENE_MONOD_{notebook}-{page_n}.xml')
        root = tree.getroot()
        page = root.find(
            '{http://schema.primaresearch.org/PAGE/gts/pagecontent/
             2013-07-15}Page')

        # Iterate over text lines detected on the page
        counter = 0
```

```
        for text_region in page.iter('{http://schema.primaresearch.org/
         PAGE/gts/pagecontent/2013-07-15}TextRegion'):
            for text_line in text_region.iter('{http://
             schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15}
             TextLine'):

                # Transkribus 'box' coords
                coords = text_line.find(
                    '{http://schema.primaresearch.org/PAGE/gts/
                     pagecontent/2013-07-15}Coords')
                coords = [double.split(',')
                          for double in coords.attrib['points'].split('
                          ')]
                coords = np.array([(int(coord[0]), int(coord[1]))
                                   for coord in coords]).astype('uint32')

                # Transkribus baseline coords
                baseline = text_line.find(
                    '{http://schema.primaresearch.org/PAGE/gts/
                     pagecontent/2013-07-15}Baseline')
                if baseline is None:
                    continue

                baseline = [double.split(
                    ',') for double in baseline.attrib['points'].split('
                     ')]
                baseline = np.array([(int(coord[0]), int(coord[1]))
                 for coord in baseline]).astype('uint32')

                # Extend the 'box'  coords
                min_y, max_y, min_x, max_x = baseline2box(
                    baseline, up=76, down=25, margin_l=18, margin_r=31)
                min_yp, max_yp, min_xp, max_xp = checkCoords(
                    min_y-pad, max_y+pad, min_x-pad, max_x+pad)

                # Select the line region
                segment = image[min_xp:max_xp, min_yp:max_yp]
                mask = np.zeros(segment.shape[:2], np.uint8)

                # Compute the line rotation angle
```

```
cv2.drawContours(
    mask, [coords - np.array([min_yp, min_xp])], -1, (
      255, 255, 255), -1, cv2.LINE_AA)
_, _, theta = cv2.minAreaRect(coords.astype('int'))
if theta > 30:
    theta = theta-90

# Compensate the rotation
rot_segment = rotate(segment, theta, cval=0)
rot_mask = rotate(mask, theta, cval=0).astype('bool')

# Compute the area of interest
try:
    min_xm, max_xm = (np.arange(rot_mask.shape[0])
      [(np.sum(rot_mask, 1)/rot_mask.shape[1]) > 0.1])[
        np.array([0, -1])]
    min_ym, max_ym = (np.arange(rot_mask.shape[1])
 [(np.sum(rot_mask, 0)/rot_mask.shape[0]) > 0,05])[
        np.array([0, -1])]
except IndexError:
    print (IndexError)
    continue

# Text for the creation of the model
text_segment = text_line.find(
        '{http://schema.primaresearch.org/PAGE/gts/
 pagecontent/2013-07-15}TextEquiv')[0].text
               text_segment = text_segment.replace('é',
 'e').replace('è', 'e').replace('ö', 'o').replace('î',
 'i').replace('ï','i').replace('ê',
 'e').replace('ù','u').replace('à','a').replace('ç','c')
 .replace('ò','o').replace('û','u').replace('ä','a')

# Text for inference
#text_segment = 'sans transcription'

# Save transcription
with open(f'{notebook}/BenthamDatasetR0-GT/
 Transcriptions/{notebook}_{page_n}_{counter}.txt',
 'w+') as f:
```

```
                    f.write(text_segment)

                # Save segment image
                cv2.imwrite(f'{notebook}/BenthamDatasetR0-GT/Images/
                 Lines/{notebook}_{page_n}_{counter}.png',
                           rot_segment[max([0, min_xm-34]):max_xm+15,
                 max([0, min_ym-18]):max_ym+10])
                list_segments.append(f'{notebook}_{page_n}_{counter}')
                counter += 1
```

```
# Create split partitions for validation, test and train

val, test = len(list_segments)*0.1, len(list_segments)*0.2
random.shuffle(list_segments)

with open(f'BenthamDatasetR0-GT/Partitions/TrainLines.lst', 'w+') as f:
    f.write(formatList(list_segments[int(val)+int(test):]))
with open(f'BenthamDatasetR0-GT/Partitions/TestLines.lst', 'w+') as f:
    f.write(formatList(list_segments[int(val):int(val)+int(test)]))
with open(f'BenthamDatasetR0-GT/Partitions/ValidationLines.lst', 'w+') as
    f:
    f.write(formatList(list_segments[:int(val)]))
```

```
# FOR INFERENCE
# Create split partitions for validation, test and train

val, test = len(list_segments)*0.1, len(list_segments)*0.2

random.shuffle(list_segments)

with open(f'BenthamDatasetR0-GT/Partitions/TrainLines.lst', 'w+') as f:
    f.write(formatList(list_segments[1:4]))
with open(f'BenthamDatasetR0-GT/Partitions/TestLines.lst', 'w+') as f:
    f.write(formatList(list_segments))
with open(f'BenthamDatasetR0-GT/Partitions/ValidationLines.lst', 'w+') as
    f:
    f.write(formatList(list_segments[4:7]))
```

ANNEXE II : ASSEMBLAGE DES PRÉDICTIONS ISSUES DE HTR-FLOR++

```
import pandas as pd
import numpy as np
import json
import glob
```

```
def getPage(filename):
    return int(filename.split('_')[1])

def getSegment(filename):
    return int(filename.split('_')[2])
```

```
cahier_n = 1

f = open(glob.glob('.../TestLines.lst'), "r")
test_lines = f.read().split('\n')[:-1]

f = open(glob.glob('.../predict_CH_ACV_P_RENE_MONOD_*.txt'), "r")
pred = f.read().split('\n')[:-1]
pred = np.array(pred)[np.arange(len(pred))%2 == 1].tolist()
pred = [p[5:] for p in pred]

df = pd.DataFrame({'lines': test_lines, 'pred': pred})

df['page'] = df['lines'].apply(getPage)
df['segment'] = df['lines'].apply(getSegment)

pages = []
for page_n in range(1, df['page'].max()+1):
    df_page = df[df["page"] == page_n].sort_values(by="segment")

    txt = ""
    for seg in df_page['pred'].values:
        txt += seg.replace('"', "'") + ' ' '\n' + ' '

    page = {
        "cahier_n": cahier_n,
```

```
        "page_n": page_n,
        "transcription": txt
    }
    pages.append(page)

with open(f'transcr_{cahier_n:03d}.json', 'w') as f:
    json.dump(pages, f)
```

ANNEXE III : DÉTECTION DES ENTITÉS QUI APPARAISSENT À TROIS REPRISES

```
# Importer la bibliothèque spacy
import spacy

# Télécharger le modèle de langue française
!python -m spacy download fr_core_news_md

# Charger le modèle de langue française
nlp = spacy.load("fr_core_news_md")
```

```
# Charger le modèle de traitement du langage naturel pour le français
nlp_fr = spacy.load('fr_core_news_lg')

# Récupérer tous les fichiers commençant par "transcr_" dans le répertoire donné
files = glob.glob('…/DATA.json')

# Initialiser un dictionnaire pour stocker les entités et leur fréquence
entity_freq = {}

# Boucler sur chaque fichier et extraire le texte de chaque document pour analyse avec Spacy
for file in files:
    with open(file, 'r') as f:
        data = json.load(f)
        for document in data:
            doc = nlp_fr(document['transcription'])
            entities = [(ent.label_, ent.text) for ent in doc.ents]
```

```
            # Ajouter chaque entité au dictionnaire et incrémenter sa
             fréquence
            for entity in entities:
                if entity in entity_freq:
                    entity_freq[entity] += 1
                else:
                    entity_freq[entity] = 1

# Filtrer les entités pour ne conserver que celles qui apparaissent au
moins trois fois
filtered_entities = [entity for entity, freq in entity_freq.items() if
freq >= 3]

# Trier les entités filtrées par ordre alphabétique
filtered_entities.sort()

# Afficher les entités filtrées avec un retour à la ligne après chaque
entité
for entity in filtered_entities:
    print(entity[1])
```

ANNEXE IV : ANALYSE DE FRÉQUENCE DES ENTITÉS NOMMÉES

```
# Importer la bibliothèque spacy
import spacy

# Télécharger le modèle de langue française
!python -m spacy download fr_core_news_md

# Charger le modèle de langue française
nlp = spacy.load("fr_core_news_md")
```

```
# Initialiser des listes vides pour stocker les entités nommées de chaque
catégorie
personnes = []
lieux = []
miscellaneous_entities = []
organisation = []
```

```
# Créer une liste d'arrêts en français
french_stopwords = ["le", "la" ,'l', "les", "de", "du", "des", "un",
"une", "et", "ou", "car", "par", "pour", "avec", "sur"]

# Ajouter des mots spécifiques à la variable 'polier'
polier_stopwords = ['mois', 'janvier', 'fevrier', 'mars', 'avril', 'mai',
'juin', 'juillet', 'aout', 'septembre', 'octobre', 'novembre',
'decembre', 'xbre', 'jour', 'jours', 'lundi', 'mardi', 'mercredi',
'jeudi', 'vendredi', 'samedi', 'dimanche', 'lun', 'mar', 'mer', 'jeu',
'ven', 'sam', `dim', 'mr', 'mrs', 'mde', 'mle', 'mlle', 'monsieur', 'mr
de']

# Concaténer les deux listes pour créer la liste complète des mots à
supprimer
stopwords = french_stopwords + polier_stopwords


# Charger le modèle de traitement du langage naturel pour le français
nlp_fr = spacy.load('fr_core_news_lg')

# Récupérer tous les fichiers commençant par "transcr_" dans le
répertoire donné
files = glob.glob(`…/DATA.json’)

# Boucler sur chaque fichier et extraire le texte de chaque document pour
analyse avec Spacy
for file in files:
    with open(file, 'r') as f:
        data = json.load(f)
        for document in data:
            doc = nlp_fr(document['transcription'])
            for ent in doc.ents:
                # Ajouter l'entité nommée à la liste appropriée en
                 fonction de sa catégorie
                if ent.label_ == 'PER' and (ent.text.lower() not in
                 stopwords):
                    personnes.append(ent.text)
                elif ent.label_ == 'LOC' and (ent.text.lower() not in
                 stopwords):
                    lieux.append(ent.text)
```

```
                elif ent.label_ == 'MISC' and (ent.text.lower() not in
                 stopwords):
                    miscellaneous_entities.append(ent.text)
                elif ent.label_ == 'ORG' and (ent.text.lower() not in
                 stopwords):
                    organisation.append(ent.text)

# Afficher les 10 entités nommées les plus fréquentes dans chaque
catégorie
print('Entités nommées les plus fréquentes:\n')
print('\nPersonnes:\n', "\n".join([f"{entity}: {count}" for entity, count
in Counter(personnes).most_common(20)]))

print('\nLieux:\n', "\n".join([f"{entity}: {count}" for entity, count in
Counter(lieux).most_common(20)]))

print('\nOrganisation:\n', "\n".join([f"{entity}: {count}" for entity,
count in Counter(organisation).most_common(20)]))

print('\nDivers:\n', "\n".join([f"{entity}: {count}" for entity, count in
Counter(miscellaneous_entities).most_common(20)]))
```

ANNEXE V : ALGORITHME DE LATENT DIRICHLET ALLOCATION

```
import spacy
import glob
import json
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
```

```
# Charger le modèle de traitement du langage naturel pour le français
nlp_fr = spacy.load('fr_core_news_lg')

# Récupérer tous les fichiers commençant par "transcr_" dans le
répertoire donné
files = glob.glob('…/DATA.json')
```

```
# Boucler sur chaque fichier et extraire le texte de chaque document pour
analyse avec Spacy
documents, tokens = [], []

for i, file in enumerate(files):
    print(f"Traitement du fichier {i+1}/{len(files)}: {file}")
    with open(file, 'r') as f:
        data = json.load(f)
        for document in data:
            doc = nlp_fr(document['transcription'])
            words = [token.lemma_ for token in doc if not token.is_stop
             and token.is_alpha and len(token.lemma_) > 2]
            tokens += words
            documents.append(words)

# Créer une liste d'arrêts en français
french_stopwords = ["le", "la" ,'l', "les", "de", "du", "des", "un",
"une", "et", "ou", "car", "par", "pour", "avec", "sur"]

# Ajouter des mots spécifiques à la variable 'polier'
polier_stopwords = ['mois', 'janvier', 'fevrier', 'mars', 'avril', 'mai',
'juin', 'juillet', 'aout', 'septembre', 'octobre', 'novembre',
'decembre', 'xbre', 'jour', 'jours', 'lundi', 'mardi', 'mercredi',
'jeudi', 'vendredi', 'samedi', 'dimanche', 'lun', 'mar', 'mer', 'jeu',
'ven', 'sam', 'dim', 'mr', 'mrs', 'mde', 'mle', 'mlle', 'monsieur', 'mr
de']

# Concaténer les deux listes pour créer la liste complète des mots à
supprimer
stopwords = french_stopwords + polier_stopwords

# Sélectionner le vocabulaire à utiliser pour LDA
vocabulaire = pd.Series(tokens).value_counts()[15:]
vocabulaire = vocabulaire[vocabulaire >= 3].sort_index().keys()

# Convertir les données en vecteurs pour LDA
count_vectorizer = CountVectorizer(stop_words=stopwords,
vocabulary=vocabulaire)
X = count_vectorizer.fit_transform([' '.join(doc) for doc in documents])
```

```
# Exécuter LDA sur les données
lda_model = LatentDirichletAllocation(n_components=15, max_iter=20,
random_state=42)
lda_model.fit(X)
```

```
# Afficher les mots les plus représentatifs de chaque cluster
feature_names = np.array(list(count_vectorizer.vocabulary_.keys()))
for topic_idx, topic in enumerate(lda_model.components_):
    top_words = [feature_names[i] for i in topic.argsort()[:-21:-1]]
    #print(f"Topic #{topic_idx}: {' `.join(top_words)}")
    print(f"Les mots les plus représentatifs du topic #{topic_idx} sont :
{' '.join(top_words)}")

print('\n')

# Mots les plus caractéristiques
type_attribution = lda_model.transform(np.identity(len(vocabulaire)))
for topic_idx, topic in enumerate(lda_model.components_):
     top_charac_words = [(vocabulaire[i], int(np.sum(X[:, i]))) for i in
np.flip(np.argsort(type_attribution[:,topic_idx]))[:20]]
    print(f"Les mots les plus caractéristiques du topic #{topic_idx} sont
: {' '.join([word[0] for word in top_charac_words])} (fréquence :
{[word[1] for word in top_charac_words]})")
```